# THÈSE de DOCTORAT

## de l'UNIVERSITÉ TOULOUSE CAPITOLE

***Présentée et soutenue par***

**Monsieur Philipp KOCH**

**Le 11 octobre 2024**

**Apprentissage automatique pour l'histoire économique**

École doctorale : **Mathématiques, Informatique et Télécommunications de Toulouse**
Spécialité : **Mathématiques et Applications**
Unité de recherche : **TSE-R - Toulouse School of Economics - Recherche**

***Thèse dirigée par*** *Monsieur Cesar HIDALGO*

***Composition du jury***

*Rapporteure : Mme Ina GANGULI*
*Rapporteur : M. Andrea MORRISON*
*Examinatrice : Mme Eva COLL-MARTINEZ*
*Examinateur : M. Balázs LENGYEL*
*Directeur de thèse : M. César A. HIDALGO*

# Contents

# ABSTRACT

Machine learning methods helped expand the economics toolbox over the past decades. Recent contributions started to explore the intersection of machine learning methods and economic history. This thesis aims to contribute to this emerging field through three chapters.

The first chapter reviews the literature and finds that applications of machine learning broadly fall into three categories: (1) ML as a pre-processing tool to digitize archival sources such as historical texts and maps, facilitating large-scale quantitative analyses; (2) unsupervised ML models, including clustering and dimensionality reduction, to derive new variables that reveal latent historical patterns and relationships; and (3) supervised ML models to generate new data or enhance existing datasets.

The second chapter showcases the use of unsupervised machine learning—that is, measures of relatedness—to help us understand why Paris became the Mecca for the arts and Vienna a beacon of classical music. Specifically, we use data on more than 22,000 historical individuals born between the years 1000 and 2000 to estimate the contribution of famous immigrants, emigrants, and locals to the knowledge specializations of European regions. We find that migrants play a crucial role in shaping European cities. The probability that a region develops or keeps specialization in an activity (based on the birth of famous physicists, painters, etc.) grows with both, the presence of immigrants with knowledge in that activity and immigrants with knowledge in related activities.

In the third chapter, we introduce a machine learning method to augment the availability of historical GDP per capita estimates. Using data on the places of birth, death, and occupations of hundreds of thousands of historical figures, we build an elastic net regression model to perform feature selection and generate out-of-sample estimates that explain 90% of the variance in known historical income levels. We use this model to generate GDP per capita estimates for dozens of countries and hundreds of regions in Europe and North America for the past 700 years. We externally validate our estimates by comparing them with four proxies of economic output and showing they reproduce the well-known reversal of fortune between southwestern and northwestern Europe between 1300 and 1800. These findings validate the use of fine-grained biographical data as a method to produce historical GDP per capita estimates.

Together, this thesis explores the potential of machine learning methods to enhance our understanding of economic history by providing a review of the state-of-the art and showcasing the use of unsupervised and supervised machine learning models to investigate questions that were left relatively unexplored.

**Keywords**: machine learning, economic history, economic complexity, network science

# RÉSUMÉ

Les méthodes d'apprentissage automatique ont contribué à élargir la boîte à outils de l'économie au cours des dernières décennies. Des contributions récentes ont commencé à explorer l'intersection des méthodes d'apprentissage automatique et de l'histoire économique. Cette thèse vise à contribuer à ce domaine émergent à travers trois chapitres.

Le premier chapitre passe en revue la littérature et constate que les applications de l'apprentissage automatique se répartissent globalement en trois catégories : (1) l'apprentissage automatique en tant qu'outil de prétraitement pour numériser les sources d'archives telles que les textes et les cartes historiques, facilitant ainsi les analyses quantitatives à grande échelle ; (2) les modèles d'apprentissage automatique non supervisés, y compris le regroupement et la réduction de la dimensionnalité, pour dériver de nouvelles variables qui révèlent des modèles et des relations historiques latents ; et (3) les modèles d'apprentissage automatique supervisés pour générer de nouvelles données ou améliorer les ensembles de données existants.

Le deuxième chapitre présente l'utilisation de l'apprentissage automatique non supervisé, c'est-à-dire des mesures de parenté, pour nous aider à comprendre pourquoi Paris est devenue la Mecque des arts et Vienne le phare de la musique classique. Plus précisément, nous utilisons des données sur plus de 22 000 individus historiques nés entre l'an 1000 et l'an 2000 pour estimer la contribution d'immigrants, d'émigrants et de locaux célèbres à la spécialisation des connaissances des régions européennes. Nous constatons que les migrants jouent un rôle crucial dans le façonnement des villes européennes. La probabilité qu'une région développe ou conserve une spécialisation dans une activité (sur la base de la naissance de physiciens, de peintres, etc. célèbres) augmente avec la présence d'immigrants possédant des connaissances dans cette activité et d'immigrants possédant des connaissances dans des activités connexes.

Dans le troisième chapitre, nous introduisons une méthode d'apprentissage automatique pour augmenter la disponibilité des estimations historiques du PIB par habitant. En utilisant des données sur les lieux de naissance, de décès et les professions de centaines de milliers de personnages historiques, nous construisons un modèle de régression à filet élastique pour effectuer une sélection des caractéristiques et générer des estimations hors échantillon qui expliquent 90 % de la variance des niveaux de revenus historiques connus. Nous utilisons ce modèle pour générer des estimations du PIB par habitant pour des dizaines de pays et des centaines de régions d'Europe et d'Amérique du Nord pour les 700 dernières années. Nous validons nos estimations en les comparant à quatre indicateurs de la production économique et en montrant qu'elles reproduisent le renversement de fortune bien connu entre le sud-ouest et le nord-ouest de l'Europe entre 1300 et 1800. Ces résultats valident l'utilisation de données biographiques fines comme méthode pour produire des estimations historiques du PIB par habitant.

Cette thèse explore le potentiel des méthodes d'apprentissage automatique pour améliorer notre compréhension de l'histoire économique en fournissant une revue de l'état de l'art et en présentant l'utilisation de modèles d'apprentissage automatique supervisés et non supervisés pour étudier des questions qui ont été laissées relativement inexplorées.

**Mots clés**: apprentissage automatique, histoire économique, complexité économique, science des réseaux

# ACKNOWLEDGEMENTS

The pursuit of a PhD, despite necessary periods of solitude, is nothing one can do completely on his own. I was lucky and privileged to meet and be supported by many inspiring people, who I would like to wholeheartedly thank for accompanying me on my PhD journey over the past three years.

First of all, I am massively grateful for the continuous and unconditional support I received from my supervisor, César A. Hidalgo. I would not be the researcher I am today if it was not for César's guidance. The amount of inspiration and ambition he conveys every single day, including each mouth-watering lunch at Café Baggio, is simply unmatched. Being his PhD student and having a first-row seat on how he thinks, develops ideas, writes, visualizes, manages etc. is a huge privilege that I am massively thankful for. And first-row seat is not meant metaphorically but literally: I still vividly remember the sessions where we sit in front of one monitor editing my latest draft word by word or arrange a full-page figure with our latest results. It is rare to have a supervisor who is as invested in his PhD student's work as César is. In fact, the German term for PhD supervisor is '*Doktorvater*', which loosely translates to '*doctor's father*'. Looking back, this feels like an appropriate analogy: Starting as an academic youngster three years ago, who just made his first scientific steps, César raised me to academic adulthood.

If I view César as my '*doctor's father*', Viktor Stojkoski must be considered my '*doctor's big brother*'. Viktor joined our group a couple of months after me as a postdoc, and somehow it immediately clicked between the three of us—not just professionally, but also personally. This becomes clear when looking at our output: In less than three years, we wrote four papers together. Besides chapters 2 and 3 of this thesis, we published a paper on Multidimensional Economic Complexity in *Communications Earth & Environment* and a paper on estimating digital trade through corporate revenue data in *Nature Communications*. It's not just that we have been massively productive as a team, but all the projects were a ton of fun, and each felt like we were onto something big. To this day, I am wondering how one can be as productive as Viktor is. Maybe having the largest biceps and the highest chicken consumption in the whole economic complexity field is the key factor. Jokes aside, it is rare to find a team working as smoothly together as we are, and I truly hope that we can keep our streak going in the coming years.

While César and Viktor were the largest professional influences during my PhD, meeting Johannes Dahlke was the best thing that happened to me on a personal level in Toulouse. Johannes visited our group during the first academic year of my PhD and both of us probably did not understand what a wonderful and deep friendship began to develop at those first, somewhat awkward dinners in Toulouse. Soon after, we moved to an apartment together and spent our nights arguing about the state of economics or AI definitions, watching twitch streams, playing billiards, and talking about anything under the sun.

There are many more people I want to thank, but, for the sake of brevity, let me conclude with a non-exhaustive list: all remaining members of the Center for Collective Learning for making it the inspiring and fun place it has been over the years, especially Lea, Carlos, Jingling, Mariana, Vieri, and Pierre-Alex; the wonderful community that formed at the PhD Summer School at Utrecht University in 2022; Jesus Crespo for pushing me during my Masters and always providing helpful comments; Chris Esposito for reaching out and inviting me to visit Chicago; all the colleagues hosting me at the Complexity Science Hub in Vienna; my colleagues at EcoAustria in Vienna; and many more.

# INTRODUCTION

During the last decades, machine learning methods helped expand the economics toolbox (1, 2), from the use of satellite images to estimate poverty (3–6), population (7, 8), and land use (9–12), to the use of recommender systems to support economic diversification policies (13–16) and the estimation of digital trade (17).

But machine learning methods are not only useful for studying the present or predicting the future, they can also be used to explore the past (18–20). This thesis aims to contribute to this emerging field at the intersection of economic history and machine learning through three chapters. The first chapter provides a review of the state-of-the-art and an outlook on promising future avenues in the field, while the other two chapters showcase the use of machine learning to better understand economic history.

What exactly do we mean by machine learning? In general, machine learning aims to optimally predict an outcome variable $y$ given inputs $\boldsymbol{X}$. While econometrics is mostly concerned with finding the data-generating process and estimating the correct functional form $\hat{f}(\boldsymbol{X})$, machine learning methods aim to find the best $\hat{y}$. Supervised machine learning models are trained on labeled input and output data (e.g. regression or classification tasks), while unsupervised machine learning models lack a predefined output which requires them to identify patterns in the data themselves (e.g. clustering or dimensionality reduction techniques). In both cases, machine learning algorithms require relatively large amounts of training data to find a mapping between input and output data or identify the underlying patterns.

Despite the frequent scarcity of large amounts of data in historical settings, applications of machine learning to economic history come in many forms. In the first chapter of this thesis, I review the literature on applications of machine learning in economic history and identify promising future research avenues. These applications broadly fall into three categories: First, machine learning is utilized as a pre-processing tool to digitize archival sources such as historical texts and maps, thus facilitating large-scale quantitative analyses. Second, unsupervised machine learning models, including clustering and dimensionality reduction, are employed to derive new variables that uncover latent historical patterns and relationships. Third, supervised machine learning models are leveraged to generate new data or enhance existing datasets. As existing methods for data pre-processing continue to advance, I argue the most promising research avenues involve exploiting machine learning models to generate novel data and adopting recent innovations to better handle unstructured text data.

The remaining two chapters of this thesis include applications of machine learning techniques using data on famous individuals (21, 22). Famous historical figures are not a capricious choice.

Accurate biographical records of famous individuals are abundant and provide one of the most comprehensive representations of historical economies, especially in preindustrial periods (23–25). Also, upper tail human capital is known to be a key driver of modern economic growth (26–28). Combined with machine learning, this data can shed light on questions that were left relatively unexplored, as the second and third chapter of this thesis show.

The second chapter (29) provides an example of using unsupervised machine learning models—that is, measures of relatedness (13, 30)—to derive new variables that help us understand the evolution of European cities over the past 1,000 years. Specifically, we ask: Did migrants make Paris a Mecca for the arts and Vienna a beacon of classical music? Or was their rise a pure consequence of local actors? We use data on more than 22,000 historical individuals born between the years 1000 and 2000 to estimate the contribution of famous immigrants, emigrants, and locals to the knowledge specializations of European regions. We find that the probability that a region develops or keeps specialization in an activity (based on the birth of famous physicists, painters, etc.) grows with the presence of immigrants with knowledge of that activity and immigrants with knowledge in related activities. In contrast, we do not find robust evidence that the presence of locals with related knowledge explains entries and/or exits. We address some endogeneity concerns using fixed-effects models considering any location-period-activity specific factors (e.g. the presence of a new university attracting scientists).

The third chapter introduces a machine learning method to augment the availability of historical GDP per capita estimates. For decades, economic historians have made great efforts to reconstruct the GDP per capita of countries and regions but estimates of historical GDPs per capita are still scarce (31, 32). This limits our ability to explore questions of long-term economic growth and development. In this chapter, we ask whether data on the biographies of hundreds of thousands of historical figures, combined with machine learning methods, can be used to extend GDP per capita estimates to countries, regions, and time periods for which this data is not available. Using data on the places of birth, death, and occupations of historical figures, we build an elastic net regression model to perform feature selection and generate out-of-sample estimates that explain 90% of the variance in known historical income levels. We use this model to generate GDP per capita estimates for dozens of countries and hundreds of regions in Europe and North America for the past 700 years. We externally validate our estimates by comparing them with four proxies of economic output: urbanization rates over the past 500 years, body height in the 18th century, wellbeing in 1850, and church building activity in the 14th and 15th century. Additionally, we show our estimates reproduce the well-known reversal of fortune between southwestern and northwestern Europe between 1300 and 1800 and find this is largely

driven by countries and regions engaged in Atlantic trade. These findings validate the use of fine-grained biographical data as a method to produce historical GDP per capita estimates.

The thesis is structured as follows. The next three chapters include the main texts of the articles outlined above, before I provide a discussion of the results in a unified context. The remaining two sections include supplementary materials for the second and third chapter, respectively.

# References


1. S. Athey, "The Impact of Machine Learning on Economics" in *The Economics of Artificial Intelligence: An Agenda*, National Bureau of Economic Research conference report., A. Agrawal, J. Gans, A. Goldfarb, National Bureau of Economic Research, Eds. (The University of Chicago Press, 2019), pp. 507–547.
2. S. Athey, G. W. Imbens, Machine Learning Methods That Economists Should Know About. *Annu. Rev. Econ.* **11**, 685–725 (2019).
3. N. Jean, *et al.*, Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
4. D. Ahn, *et al.*, A human-machine collaborative approach measures economic development using satellite imagery. *Nat. Commun.* **14**, 6811 (2023).
5. G. Chi, H. Fang, S. Chatterjee, J. E. Blumenstock, Microestimates of wealth for all low- and middle-income countries. *Proc. Natl. Acad. Sci.* **119**, e2113658119 (2022).
6. J. V. Henderson, A. Storeygard, D. N. Weil, Measuring Economic Growth from Outer Space. *Am. Econ. Rev.* **102**, 994–1028 (2012).
7. C. Robinson, F. Hohman, B. Dilkina, A Deep Learning Approach for Population Estimation from Satellite Imagery in *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, (ACM, 2017), pp. 47–54.
8. F. R. Stevens, A. E. Gaughan, C. Linard, A. J. Tatem, Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLOS ONE* **10**, e0107042 (2015).
9. M. C. Hansen, *et al.*, High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* **342**, 850–853 (2013).
10. E. Rolf, *et al.*, A generalizable and accessible approach to machine learning with global satellite imagery. *Nat. Commun.* **12**, 4392 (2021).
11. M. Burke, A. Driscoll, D. B. Lobell, S. Ermon, Using satellite imagery to understand and promote sustainable development. *Science* **371**, eabe8628 (2021).
12. L. Yu, *et al.*, Meta-discoveries from a synthesis of satellite-based land-cover mapping research. *Int. J. Remote Sens.* **35**, 4573–4588 (2014).
13. C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann, The product space conditions the development of nations. *Science* **317**, 482–487 (2007).
14. F. L. Pinheiro, D. Hartmann, R. Boschma, C. A. Hidalgo, The time and frequency of unrelated diversification. *Res. Policy* 104323 (2021). https://doi.org/10.1016/j.respol.2021.104323.
15. S. Poncet, F. S. de Waldemar, Product Relatedness and Firm Exports in China. *World Bank Econ. Rev.* **29**, 579–605 (2015).

16. C. A. Hidalgo, Economic complexity theory and applications. *Nat. Rev. Phys.* **3**, 92–113 (2021).
17. V. Stojkoski, P. Koch, E. Coll, C. A. Hidalgo, Estimating digital product trade through corporate revenue data. *Nat. Commun.* **15**, 5262 (2024).
18. P.-P. Combes, L. Gobillon, Y. Zylberberg, Urban economics in a historical perspective: Recovering data with machine learning. *Reg. Sci. Urban Econ.* **94**, 103711 (2022).
19. W. W. Hanlon, S. Heblich, History and urban economics. *Reg. Sci. Urban Econ.* **94**, 103751 (2022).
20. M. P. Gutmann, E. K. Merchant, E. Roberts, "Big Data" in Economic History. *J. Econ. Hist.* **78**, 268–299 (2018).
21. A. Z. Yu, S. Ronen, K. Hu, T. Lu, C. A. Hidalgo, Pantheon 1.0, a manually verified dataset of globally famous biographies. *Sci. Data* **3**, 150075 (2016).
22. M. Laouenan, *et al.*, A cross-verified database of notable people, 3500BC-2018AD. *Sci. Data* **9**, 290 (2022).
23. M. Schich, *et al.*, A network framework of cultural history. *Science* **345**, 558–562 (2014).
24. M. Serafinelli, G. Tabellini, Creativity over time and space: A historical analysis of European cities. *J. Econ. Growth* **27**, 1–43 (2022).
25. D. De La Croix, O. Licandro, The longevity of famous people from Hammurabi to Einstein. *J. Econ. Growth* **20**, 263–303 (2015).
26. J. Mokyr, The Intellectual Origins of Modern Economic Growth. *J. Econ. Hist.* **65** (2005).
27. J. Mokyr, *A Culture of Growth: The Origins of the Modern Economy* (Princeton University Press, 2017).
28. J. Mokyr, H.-J. Voth, Understanding Growth in Europe, 1700-1870: Theory and Evidence. *Camb. Econ. Hist. Mod. Eur.* **1**, 7–42 (2010).
29. P. Koch, V. Stojkoski, C. A. Hidalgo, The Role of Immigrants, Emigrants, and Locals in the Historical Formation of European Knowledge Agglomerations. *Reg. Stud.* 1–15 (2023). https://doi.org/10.1080/00343404.2023.2275571.
30. C. A. Hidalgo, *et al.*, "The Principle of Relatedness" in *Unifying Themes in Complex Systems IX*, Springer Proceedings in Complexity., A. J. Morales, C. Gershenson, D. Braha, A. A. Minai, Y. Bar-Yam, Eds. (Springer International Publishing, 2018), pp. 451–457.
31. J. Bolt, J. L. van Zanden, Maddison style estimates of the evolution of the world economy. A new 2020 update. *Maddison-Proj. Work. Pap.* **WP-15** (2020).
32. J. Bolt, J. L. van Zanden, The Maddison Project: collaborative research on historical national accounts: The Maddison Project. *Econ. Hist. Rev.* **67**, 627–651 (2014).

# CHAPTER 1:
# TOWARDS USING MACHINE LEARNING TO BETTER UNDERSTAND ECONOMIC HISTORY

# Towards using machine learning to better understand economic history

*Philipp Koch*[1,2]

[1] Center for Collective Learning, ANITI, IRIT, Université de Toulouse, Toulouse, France.
[2] EcoAustria – Institute for Economic Research, Vienna, Austria.

**Abstract**

Machine learning (ML) methods are transforming quantitative research across fields. Here, I review the emerging research field at the intersection of ML and economic history and provide outlooks on future research avenues. Applications of ML in economic history broadly fall into three categories: First, ML is utilized as a pre-processing tool to digitize archival sources such as historical texts and maps, thus facilitating large-scale quantitative analyses. Second, unsupervised ML models, including clustering and dimensionality reduction, are employed to derive new variables that uncover latent historical patterns and relationships. Third, supervised ML models are leveraged to generate new data or enhance existing datasets. As existing methods for data pre-processing continue to advance, I argue the most promising research avenues at the intersection of economic history and ML involve exploiting ML models to generate novel data and adopting recent innovations in handling unstructured text data such as Large Language Models (LLMs).

# Introduction

Hundreds of papyrus scrolls were buried in the villa of Julius Caesar's father-in-law after the eruption of Mount Vesuvius in 79 AD. Almost 2,000 years later, a team of three young engineers deciphered parts of one scroll, building on detailed 3D scans of the heavily charred document.[*] Using machine learning techniques, they could exploit marginal differences in the thickness of the scanned scroll to identify ink remains, eventually recovering writings that were thought to be lost.

While it is enlightening to read those contemplations of a Roman intellectual on whether scarce food brings more pleasure, this breakthrough showcases something more fundamental: the ability of machine learning to enhance our understanding of the past. Here, I give an overview of the emerging research field at the intersection of machine learning and economic history, and provide an outlook on where I believe promising research avenues open.

Indeed, machine learning methods are becoming more and more prevalent in the field of economics (Athey, 2019; Athey et al., 2021; Athey & Imbens, 2019). Applications range from the use of satellite images to estimate economic prosperity (Ahn et al., 2023; Chi et al., 2022; Henderson et al., 2012; Jean et al., 2016), to the use of recommender systems and dimensionality reduction techniques to inform economic policy (Hidalgo, 2021; Hidalgo et al., 2007; Pinheiro et al., 2021; Poncet & de Waldemar, 2015). As some articles argued (Combes et al., 2022; Gutmann et al., 2018; Hanlon & Heblich, 2022), and as we will see in this review, machine learning is increasingly used to explore historical research questions and has the potential to lift our understanding of the past to another level.

But what exactly do we understand as machine learning? In general, machine learning (ML) aims to optimally predict an outcome variable $y$ given inputs $\boldsymbol{X}$. While econometrics is mostly concerned with finding the data-generating process and estimating the correct functional form $\hat{f}(\boldsymbol{X})$, ML methods aim to find the best $\hat{y}$. Supervised ML models are trained on labeled input and output data (e.g. regression or classification tasks), while unsupervised ML models lack a predefined output which requires them to identify patterns in the data themselves (e.g. clustering or dimensionality reduction techniques). In both cases, ML algorithms require relatively large amounts of training data that allow them to find a mapping between input and output data or identify the underlying patterns.

[*] https://scrollprize.org/grandprize

As I will argue in this review, the use of ML in economic history can be broadly categorized into three clusters: (1) ML as a pre-processing tool, (2) unsupervised ML to derive new insightful variables, and (3) supervised ML to generate new data (Fig. 1). Currently, the most prevalent use of ML in economic history is exploiting its capacity as a pre-processing tool to enable the quantitative analysis of historical sources. This involves, for instance, text recognition or probabilistic matching of observations across datasets. While this will certainly remain a fruitful field of research due to a continuous improvement of the underlying methods, the frontier in the intersection between economic history and machine learning involves methods to derive new insightful variables through unsupervised ML methods, or even to generate novel data exploiting the predictive power of supervised ML methods. Also, recent advances in how ML and artificial intelligence models handle text, such as Large Language Models, can have a significant impact on the field of economic history.

This review is structured as follows. First, I will describe how machine learning is used as a pre-processing tool to digitize archival sources. Then, I outline several contributions using unsupervised ML models to derive new variables, before I present recent articles using supervised ML models to generate new data. Lastly, I conclude and provide potential avenues I believe research at the intersection of machine learning and economic history is moving towards.

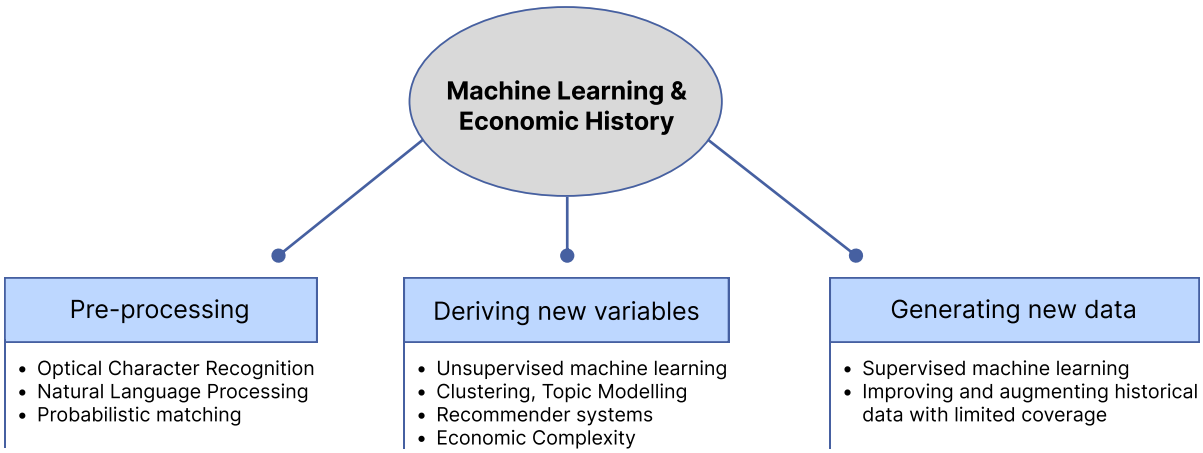



***Figure 1**. Summary of functions and methods at the intersection of machine learning and economic history.*

## Machine learning as a pre-processing tool

Natural language processing and optical character recognition (OCR) provided great advances over the past decades (Bailey et al., 2019; Gentzkow et al., 2019; Hirschberg & Manning, 2015).

Research in economic history profited from this progress and started to use ML as a pre-processing tool for quantitative analyses. For instance, digitizing archival or printed sources, a key aspect of research in economic history, scales with potent ML algorithms. This does not only apply to text but also to maps and georeferenced information. Besides transforming printed sources into machine-readable data, linking datasets such as multiple waves of census waves is a frequent use case. In this chapter, I provide an overview of how ML is used as a pre-processing tool in the field of economic history.

**Digitizing text and maps**

Text and optical character recognition algorithms allow to make printed sources machine-readable. Take historical patents, a valuable resource for innovation research (Andrews, 2021). These documents contain lots of information, including the names of inventors, some sociodemographic data, their location, and detailed descriptions of the technology itself. Recent contributions made this information accessible by using neural networks to retrieve geographical information from the scans of patents in the United States between 1836 and 1975 (Petralia et al., 2016b). All gathered information is publicly available in the HistPat dataset (Petralia et al., 2016a) and can be merged with recent patent data. Having long and consistent time-series of innovation activities is a fruitful resource. Research building on HistPat shows, for instance, that the most complex technologies increasingly concentrate in large cities since the year 1850 (Balland et al., 2020).

Van Der Wouden (2020) built upon the HistPat database and used supervised ML models to identify the names of all inventors on historical patents, while HistPat only reports the first inventor mentioned. The results indicate that inventor teams grew remarkably after 1930, mostly driven by within-city collaborations. Similarly, patents contain information on the inventors' place of residence and their country of origin. This information can be exploited to identify and analyze technologies patented by immigrants to the United States (Diodato et al., 2022). Using algorithms for data processing analogous to Petralia et al. (2016b), Diodato et al. (2022) find that the patenting activity of natives increases with the number of patents by immigrants. Also, the appearance of a new technological class correlates positively with the inventive activity of immigrants, supporting the notion of migrants as agents of structural change and a crowding-in effect of migration in innovation.

There are plenty of other historical text sources that ML techniques help make available in a structured way. Nedelkoska et al. (2021) leverage OCR and classification models based on neural networks to digitize the U.S. Dictionary of Occupational Titles (DOT) going back to

1939 and link this data with more recent occupation classifications. This allows them to study the effect of technology on the gender pay gap over the course of 80 years, finding that computerization contributed to widening the gender pay gap. Juhász & Steinwender (2018) use OCR to digitize customs records and shipping information for ports in 19th century London to explore the role of the telegraph on cotton trade. They find that the telegraph led to a larger increase of intermediate products compared to final products, because the technology allowed for transmitting more detailed information on the intermediate products' properties. Crucially, recent methodological advances allow for digitizing more complex layouts such as the Japanese language. Shen et al. (2020) use deep learning to digitize more than 2,000 historical documents describing 50,000 famous Japanese individuals. Other examples for digitized historical text sources are newspapers (Bingham, 2010), city directories describing granular geographic data (Albers & Kappner, 2023), and cultural heritage institutions such as museums and libraries more generally (Sporleder, 2010).

But text sources are not the only printed historical sources that ML helps digitize. Another valuable source of information are historical maps. Combes et al. (2022) describe a random forest and neural network technique to classify printed maps of 19th century France, while Flückiger et al. (2022) use the digitized Barrington Atlas and geoinformation on Roman ceramics to recreate the Roman transport network and show that these relations still matter today. Persistent path-dependency is also visible when investigating neighborhood sorting. Heblich et al. (2021) use clustering algorithms to geo-locate 19th century census entries based on a fraction of well-matched individuals and combine this information with the geo-location of 5,000 industrial chimneys. They find that the atmospheric pollution caused by these chimneys still affects neighborhood sorting and segregation in English cities today.

**Linking datasets and observations**

A key data source in economic history is census data. Census data provides highly granular information on sociodemographic characteristics of individuals, both cross-sectionally and longitudinally. To describe longitudinal relationships, it is essential to link datasets and observations across census waves. But this is far from trivial, since census data lack unique and consistent identifiers. Recent efforts, hence, propose to use basic sociodemographic information such as name, age, and gender to link observations across time (Abramitzky et al., 2021). Similar obstacles are present when linking firms across waves of manufacturing census (Hornbeck et al., 2023) or when linking family member across generations (Price et al., 2021). These are challenges where machine learning can significantly help.

Abramitzky et al. (2021) provide an overview of different census matching algorithms and compare their respective performance. Specifically, they compare automated rule-based matching algorithms without ML (Abramitzky et al., 2012, 2014), supervised ML algorithms (Feigenbaum, 2016), unsupervised ML algorithms (Abramitzky et al., 2020), and manual linking. They find that manual linking and supervised ML algorithms provide the highest matching rates, i.e. the largest share of correct matches of all observations, but with a substantial amount of false positives. In contrast, unsupervised ML algorithms exhibit the lowest matching rate with almost no false positives. Automated rule-based matching algorithms lie in-between. Given this trade-off, there is no clear recommendation on which method to use. Still, some contributions find that a significant number of false positives induced by machine learning algorithms can cause biases in empirical research building on it (Bailey et al., 2019). That is, reducing false positives might be more important than increasing matching rates.

## Unsupervised machine learning to derive new variables

ML does not just help as a pre-processing tool for digitizing archival sources. ML techniques can also be used to derive new variables describing latent and complex relationships traditional methods might miss. Especially unsupervised ML models such as clustering algorithms and dimensionality reduction techniques can be helpful in that context. Applying these new variables to economic models and regressions, or even exploring these variables descriptively, can help enhance our understanding of economic history.

Take structural topic modelling (STM), a ML technique that identifies topics within large amounts of text by taking document metadata into account to improve the estimation and interpretation of topic prevalence (Roberts et al., 2013). Grajzl and Murrell (2021a, 2021b) apply STM to understand the evolution of English caselaw and legal ideas from 1550 to 1764. The authors identify 100 distinct legal topics, such as financial claims, property law, precedence, or bankruptcy. Exploring the prevalence of topics across time, they find that, for instance, cases revolving around financial claims peaked in the late 17$^{th}$ century and that contract cases were more prevalent in the 16$^{th}$ and 17$^{th}$ century than the 18$^{th}$ century.

A methodologically similar approach of dimensionality reduction was taken by Turchin et al. (2018) to identify nine "complexity characteristics" that describe the evolution of human societies across thousands of years. After systematically collecting 51 variables from various historical and archaeological sources, they used Principal Component Analysis (PCA) to find nine meaningful clusters of variables, from the forms of government to the information systems a society used.

A dimensionality reduction technique that can explain cross-country differences in economic growth (Hidalgo & Hausmann, 2009; Koch, 2021; Stojkoski et al., 2016, 2023), income inequality (Hartmann et al., 2017; Stojkoski et al., 2023), and emissions (Romero & Gramkow, 2021; Stojkoski et al., 2023) is economic complexity. Specifically, economic complexity identifies the factors that best explain the geography of economic activities (exports, patents, research etc.) and can be understood as a proxy of productive or innovative capabilities (Hidalgo, 2021; Hidalgo & Hausmann, 2009). Data on the historical geography of economic activities, however, is scarce. Still, two contributions use the concept of economic complexity in a historical setting. Weber et al. (2021) use colonial statistics to collect disaggregated trade data for 1897 to 1906 and find that economic complexity predicts income levels 100 years later. Domini (2022) applies the concept of economic complexity to data on universal exhibitions held in Paris between 1850 and 1900, where countries could showcase their export products. Both today's income levels and economic growth in the past century are positively correlated with economic complexity.

Another measure that can be retrieved from the geography of economic activities is relatedness (Hidalgo et al., 2007). Measures of relatedness quantify how far a location is from a specific economic activity, building on the proximity between activities derived from their co-occurrences in a location. These measures have been shown to be robust predictors of entries to new activities: The more related a location is to an activity, the more likely this location is to develop a specialization in it (Hidalgo et al., 2018). In the context of machine learning, relatedness measures belong to the class of recommender systems (Lü et al., 2012).

While historical data on the spatial distribution of economic activities is scarce, information on famous individuals including their places of birth, places of death, and occupation are abundant. Hence, biographies of famous individuals can be used to derive measures of relatedness. Specifically, Koch et al. (2023) use data on thousands of famous individuals living in Europe over the past 1,000 years to create separate measures of relatedness for immigrants, emigrants, and locals. Armed with these measures, they explore how migrants and locals shape the evolution of regional specializations. They find that migrants are drivers of structural change both within the same activity and across activities. Put differently, the results show that the immigration of famous mathematicians does not only help a region give birth to famous mathematicians in the future, but also helps a region give birth to famous individuals in related fields (physics, chemistry etc.).

## Supervised machine learning to generate new data

At the core of machine learning's competences is the ability to make predictions. That is, identifying robust relationships between input and outcome variables, and predicting outcomes with new inputs. As several contributions I discuss in this section show, this ability allows us to generate and estimate new data with the help of supervised ML models.

Consider data on preindustrial economic development. Despite significant efforts of economic historians to reconstruct historical GDP per capita estimates, their availability is still limited. In a recent article, Koch et al. (2024) exploit data on hundreds of thousands of famous individuals combined with supervised ML models to augment the availability of historical GDP per capita for countries and regions in Europe and North America since the year 1300. Specifically, they use elastic net regression models with promising out-of-sample performance measures and validate their estimates by finding a high correlation with available proxies of economic development such as urbanization or body height. Similarly, data from historical Islamic biographies can be used to improve and extend preindustrial city-level population estimates (Chaney, 2022).

Just as historical GDP per capita data did not remain in our collective memory, we do not remember the location of some cities of the Bronze Age in today's Turkey, Iraq, and Syria (Barjamovic et al., 2019). But information about trade between those cities was preserved. Specifically, merchants kept record of commercial transactions on clay tablets. Those 4,000 years old tablets were recovered at archeological sites in Turkey and digitized (Barjamovic, 2011). Today we know that trade follows gravity-like patterns (Anderson, 2011). That is, simply put, larger cities trade more with each other, and trade decreases with distance. Transferring this knowledge to the past, gravity models can be trained with the locations and trade relationships of the few known Bronze Age cities to predict the location of lost cities (Barjamovic et al., 2019). In fact, the authors find that the quantitatively recovered city locations follow qualitative evidence by historians.

Limited data coverage also applies to occupational income scores in the US census. These describe average earnings for each occupation but do not consider demographics, industry, or geography, and are not available prior to 1950. Recent efforts (Saavedra & Twinam, 2020) use ML techniques to construct adjusted occupational income scores using information on industry, occupation, geography, and demographics. This improves estimates of race and gender pay gaps, and extends coverage back to 1850 (Saavedra & Twinam, 2020).

Lastly, ML models can be used to analyze other products of human culture such as artworks. Art changes massively over time, and so does the portrayal of humans within paintings. Machine learning models can learn the relationship between facial cues and human perceptions such as social trustworthiness from labeled training data and apply this to historical artworks. Investigating thousands of portraits of the National Portraits Gallery and the Web Gallery of Art, Safra et al. (2020) find that trustworthiness, measured based on clues that are today associated with trust, increased over the past 500 years and correlates with living standards.

# Discussion

In the past decade, ML became a crucial methodological resource in economics (Athey, 2019; Athey et al., 2021; Athey & Imbens, 2019). As we saw in this review, ML is starting to become an asset in economic history as well.

The ML applications in economic history research broadly fall into three clusters. First, text recognition and natural language processing are applied to digitize archival resources and make them available for empirical research. This ranges from patents (Diodato et al., 2022; Petralia et al., 2016b; Van Der Wouden, 2020) to occupation classifications (Nedelkoska et al., 2021) and census data (Abramitzky et al., 2021; Hornbeck et al., 2023; Price et al., 2021). Second, unsupervised machine learning models are used to derive new variables that help us understand historical developments. This involves text-based methods such as structural topic modelling (Grajzl & Murrell, 2021a, 2021b), but also dimensionality reduction techniques such as economic complexity (Domini, 2022; Weber et al., 2021) and recommender systems such as relatedness measures (Koch et al., 2023). Third, recent efforts generate new data building upon supervised machine learning models. While some contributions focus on augmenting the availability of historical data such as GDP per capita estimates (Koch et al., 2024), population (Chaney, 2022), or occupational income scores (Saavedra & Twinam, 2020), others learn locations of lost ancient cities from trade relationships (Barjamovic et al., 2019).

These starting points offer a variety of promising avenues for future research.

First, the use of ML in academic research is still at the beginning. Although ML is widely used todays, adoption rates soared only after 2015 (Duede et al., 2024; Gao & Wang, 2024). Hence, we can expect ML methods for pre-processing data to keep improving. For instance, ML matching algorithms still produce a significant amount of false positives, which can bias results in empirical research (Bailey et al., 2019). More generally, I believe that ML methods will become more readily available to researchers, and more reliable. Also, I see potential in moving from pre-processing text towards pre-processing multiple modes of data sources. Consider the

large amounts of audio, video or photo records that humanity has gathered over the past century. Multimodal ML methods that process and combine different types of data can be powerful tools.

Second, the frontier in ML and artificial intelligence methods will impact research in economics and economic history. Consider Large Language Models (LLMs) that took the world by storm after 2022. LLMs enable a completely different approach to handling large amounts of text data. Take data on famous individuals. In recent years, several contributions used biographies of famous individuals, e.g. extracted from Wikipedia, to analyze migration patterns or describe the historical geography of knowledge more generally (De La Croix & Licandro, 2015; Koch et al., 2023, 2024; Laouenan et al., 2022; Mokyr, 2005; Schich et al., 2014; Serafinelli & Tabellini, 2022). In general, I believe that biographies of famous individuals are a promising data resource, since they are—despite their shortcomings—one of the most comprehensive representations of historical economies. But the individuals' migration patterns could only be approximated using their place of birth and place of death. While this is a solid proxy (Koch et al., 2023), famous individuals have been remarkably mobile (Mokyr, 2005). Einstein was born in Ulm in Germany and died in Princeton but lived in several cities in the German-speaking world in the meantime. All this information is available in encyclopedias as unstructured text data, where recent advances such as LLMs can help extract information in a structured manner.

Third, I believe that the use of ML models to generate new data or improve existing data is the most promising future avenue of research. Recovering the location of lost cities (Barjamovic et al., 2019) or augmenting the availability of historical GDP per capita estimates (Koch et al., 2024) are important milestones in better understanding the past. The impact of these approaches, however, crucially depends on the quantity and quality of the available data. With better ML methods for pre-processing and the adoption of new technologies such as LLMs, the quantity and quality of input data will increase substantially.

Together, this review showed that ML is transforming research in economic history in several ways and helps us better understand certain aspects of the past. But just as only a small fraction of the papyrus scrolls found in Julius Caesar's mansion have been deciphered up to now, I believe that we are still at the beginning of seeing ML techniques impact economic history research.

# References


Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., & Pérez, S. (2021). Automated Linking of Historical Data. *Journal of Economic Literature*, *59*(3), 865–918. https://doi.org/10.1257/jel.20201599

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2012). Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. *American Economic Review*, *102*(5), 1832–1856. https://doi.org/10.1257/aer.102.5.1832

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2014). A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration. *Journal of Political Economy*, *122*(3), 467–506. https://doi.org/10.1086/675805

Abramitzky, R., Mill, R., & Pérez, S. (2020). Linking individuals across historical sources: A fully automated approach*. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *53*(2), 94–111. https://doi.org/10.1080/01615440.2018.1543034

Ahn, D., Yang, J., Cha, M., Yang, H., Kim, J., Park, S., Han, S., Lee, E., Lee, S., & Park, S. (2023). A human-machine collaborative approach measures economic development using satellite imagery. *Nature Communications*, *14*(1), 6811. https://doi.org/10.1038/s41467-023-42122-8

Albers, T. N. H., & Kappner, K. (2023). Perks and pitfalls of city directories as a micro-geographic data source. *Explorations in Economic History*, *87*, 101476. https://doi.org/10.1016/j.eeh.2022.101476

Anderson, J. E. (2011). The Gravity Model. *Annual Review of Economics*, *3*(1), 133–160. https://doi.org/10.1146/annurev-economics-111809-125114

Andrews, M. J. (2021). Historical patent data: A practitioner's guide. *Journal of Economics & Management Strategy*, *30*(2), 368–397. https://doi.org/10.1111/jems.12414

Athey, S. (2019). The Impact of Machine Learning on Economics. In A. Agrawal, J. Gans, A. Goldfarb, & National Bureau of Economic Research (Eds.), *The economics of artificial intelligence: An agenda* (pp. 507–547). The University of Chicago Press.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2021). Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association*, *116*(536), 1716–1730. https://doi.org/10.1080/01621459.2021.1891924

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, *11*(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Bailey, M., Cole, C., Henderson, M., & Massey, C. (2019). *How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data* (w24019; p. w24019). National Bureau of Economic Research. https://doi.org/10.3386/w24019

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., & Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature Human Behaviour*, *4*(3), 248–254. https://doi.org/10.1038/s41562-019-0803-3

Barjamovic, G. (2011). *A historical geography of Anatolia in the Old Assyrian colony period* (Vol. 38). Museum Tusculanum Press.

Barjamovic, G., Chaney, T., Coşar, K., & Hortaçsu, A. (2019). Trade, Merchants, and the Lost Cities of the Bronze Age. *The Quarterly Journal of Economics*, *134*(3), 1455–1503. https://doi.org/10.1093/qje/qjz009

Bingham, A. (2010). "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians." *Twentieth Century British History*, *21*(2), 225–231. https://doi.org/10.1093/tcbh/hwq007

Chaney, E. (2022). *Modern Library Holdings and Historic City Growth*. https://www.tse-fr.eu/sites/default/files/TSE/documents/sem2022/bid/chaney.pdf

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, *119*(3), e2113658119. https://doi.org/10.1073/pnas.2113658119

Combes, P.-P., Gobillon, L., & Zylberberg, Y. (2022). Urban economics in a historical perspective: Recovering data with machine learning. *Regional Science and Urban Economics*, *94*, 103711. https://doi.org/10.1016/j.regsciurbeco.2021.103711

De La Croix, D., & Licandro, O. (2015). The longevity of famous people from Hammurabi to Einstein. *Journal of Economic Growth*, *20*(3), 263–303. https://doi.org/10.1007/s10887-015-9117-0

Diodato, D., Morrison, A., & Petralia, S. (2022). Migration and invention in the Age of Mass Migration. *Journal of Economic Geography*, *22*(2), 477–498. https://doi.org/10.1093/jeg/lbab032

Domini, G. (2022). Patterns of specialization and economic complexity through the lens of universal exhibitions, 1855-1900. *Explorations in Economic History*, 101421. https://doi.org/10.1016/j.eeh.2021.101421

Duede, E., Dolan, W., Bauer, A., Foster, I., & Lakhani, K. (2024). *Oil & Water? Diffusion of AI Within and Across Scientific Fields* (arXiv:2405.15828). arXiv. http://arxiv.org/abs/2405.15828

Feigenbaum, J. (2016). *A Machine Learning Approach to Census Record Linking*. Harvard repository. https://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaum-censuslink.pdf

Flückiger, M., Hornung, E., Larch, M., Ludwig, M., & Mees, A. (2022). Roman Transport Network Connectivity and Economic Integration. *The Review of Economic Studies*, *89*(2), 774–810. https://doi.org/10.1093/restud/rdab036

Gao, J., & Wang, D. (2024). *Quantifying the Benefit of Artificial Intelligence for Scientific Research* (arXiv:2304.10578). arXiv. http://arxiv.org/abs/2304.10578

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Grajzl, P., & Murrell, P. (2021a). A machine-learning history of English caselaw and legal ideas prior to the Industrial Revolution I: Generating and interpreting the estimates. *Journal of Institutional Economics*, *17*(1), 1–19. https://doi.org/10.1017/S1744137420000326

Grajzl, P., & Murrell, P. (2021b). A machine-learning history of English caselaw and legal ideas prior to the Industrial Revolution II: Applications. *Journal of Institutional Economics*, *17*(2), 201–216. https://doi.org/10.1017/S1744137420000363

Gutmann, M. P., Merchant, E. K., & Roberts, E. (2018). "Big Data" in Economic History. *The Journal of Economic History*, *78*(1), 268–299. https://doi.org/10.1017/S0022050718000177

Hanlon, W. W., & Heblich, S. (2022). History and urban economics. *Regional Science and Urban Economics*, *94*, 103751. https://doi.org/10.1016/j.regsciurbeco.2021.103751

Hartmann, D., Guevara, M. R., Jara-Figueroa, C., Aristarán, M., & Hidalgo, C. A. (2017). Linking Economic Complexity, Institutions, and Income Inequality. *World Development*, *93*, 75–93. https://doi.org/10.1016/j.worlddev.2016.12.020

Heblich, S., Trew, A., & Zylberberg, Y. (2021). East-Side Story: Historical Pollution and Persistent Neighborhood Sorting. *Journal of Political Economy*, *129*(5), 1508–1552. https://doi.org/10.1086/713101

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring Economic Growth from Outer Space. *American Economic Review*, *102*(2), 994–1028. https://doi.org/10.1257/aer.102.2.994

Hidalgo, C. A. (2021). Economic complexity theory and applications. *Nature Reviews Physics*, *3*, 92–113. https://doi.org/10.1038/s42254-020-00275-1

Hidalgo, C. A., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., Glaeser, E., He, C., Kogler, D. F., Morrison, A., Neffke, F., Rigby, D., Stern, S., Zheng, S., & Zhu, S. (2018). The Principle of Relatedness. In A. J. Morales, C. Gershenson, D. Braha, A. A. Minai, & Y. Bar-Yam (Eds.), *Unifying Themes in Complex Systems IX* (pp. 451–457). Springer International Publishing. https://doi.org/10.1007/978-3-319-96661-8_46

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *PNAS*, *106*(26), 10570–10575.

Hidalgo, C. A., Klinger, B., Barabási, A.-L., & Hausmann, R. (2007). The product space conditions the development of nations. *Science*, *317*(5837), 482–487. https://doi.org/10.1126/science.1144581

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hornbeck, R., Hsu, S. H.-M., Humlum, A., & Rotemberg, M. (2023). *Technological Stickiness: Switching and Entry in the Long Transition from Water to Steam Power*. https://www.stern.nyu.edu/sites/default/files/2023-10/mrotembreg_nyu_macro_steam.pdf

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Juhász, R., & Steinwender, C. (2018). *Spinning the Web: The Impact of ICT on Trade in Intermediates and Technology Diffusion* (w24590; p. w24590). National Bureau of Economic Research. https://doi.org/10.3386/w24590

Koch, P. (2021). Economic complexity and growth: Can value-added exports better explain the link? *Economics Letters*, *198*, 109682. https://doi.org/10.1016/j.econlet.2020.109682

Koch, P., Stojkoski, V., & Hidalgo, C. A. (2023). The Role of Immigrants, Emigrants, and Locals in the Historical Formation of European Knowledge Agglomerations. *Regional Studies*, 1–15. https://doi.org/10.1080/00343404.2023.2275571

Koch, P., Stojkoski, V., & Hidalgo, C. A. (2024). *Augmenting the availability of historical GDP per capita estimates through machine learning*. kochphillipp.com. https://static1.squarespace.com/static/646fc00fbb5b0c3b5be4f496/t/6684044a1fcabb0c6d1e9889/1719927888376/manuscript_historicalGDPpc_20240702.pdf

Laouenan, M., Bhargava, P., Eyméoud, J.-B., Gergaud, O., Plique, G., & Wasmer, E. (2022). A cross-verified database of notable people, 3500BC-2018AD. *Scientific Data*, *9*(1), 290. https://doi.org/10.1038/s41597-022-01369-4

Lü, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K., & Zhou, T. (2012). Recommender systems. *Physics Reports*, *519*(1), 1–49. https://doi.org/10.1016/j.physrep.2012.02.006

Mokyr, J. (2005). Mobility, Creativity, and Technological Development: David Hume, Immanuel Kant and the Economic Development of Europe. *Session on "Creativity and the Economy", German Association of Philosophy*. https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=CF581541DADAFCD43FED5624F5F3BAE3?doi=10.1.1.524.4828&rep=rep1&type=pdf

Nedelkoska, L., Matha, S. G., McNerney, J., Assumpcao, A., Diodato, D., & Neffke, F. (2021). *Eight Decades of Changes in Occupational Tasks, Computerization and the Gender Pay Gap*. https://www.ifo.de/sites/default/files/events/2022/pillars22-Nedelkoska.pdf

Petralia, S., Balland, P.-A., & Rigby, D. (2016a). *HistPat Dataset* [dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/BPC15W

Petralia, S., Balland, P.-A., & Rigby, D. L. (2016b). Unveiling the geography of historical patents in the United States from 1836 to 1975. *Scientific Data*, *3*(1), 160074. https://doi.org/10.1038/sdata.2016.74

Pinheiro, F. L., Hartmann, D., Boschma, R., & Hidalgo, C. A. (2021). The time and frequency of unrelated diversification. *Research Policy*, 104323. https://doi.org/10.1016/j.respol.2021.104323

Poncet, S., & de Waldemar, F. S. (2015). Product Relatedness and Firm Exports in China. *The World Bank Economic Review*, *29*(3), 579–605. https://doi.org/10.1093/wber/lht037

Price, J., Buckles, K., Van Leeuwen, J., & Riley, I. (2021). Combining family history and machine learning to link historical records: The Census Tree data set. *Explorations in Economic History*, *80*, 101391. https://doi.org/10.1016/j.eeh.2021.101391

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, *4*(1), 1–20.

Romero, J. P., & Gramkow, C. (2021). Economic complexity and greenhouse gas emissions. *World Development*, *139*, 105317. https://doi.org/10.1016/j.worlddev.2020.105317

Saavedra, M., & Twinam, T. (2020). A machine learning approach to improving occupational income scores. *Explorations in Economic History*, *75*, 101304. https://doi.org/10.1016/j.eeh.2019.101304

Safra, L., Chevallier, C., Grèzes, J., & Baumard, N. (2020). Tracking historical changes in perceived trustworthiness in Western Europe using machine learning analyses of facial cues in paintings. *Nature Communications*, *11*(1), 4728. https://doi.org/10.1038/s41467-020-18566-7

Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., & Helbing, D. (2014). A network framework of cultural history. *Science*, *345*(6196), 558–562. https://doi.org/10.1126/science.1240064

Serafinelli, M., & Tabellini, G. (2022). Creativity over time and space: A historical analysis of European cities. *Journal of Economic Growth*, *27*(1), 1–43. https://doi.org/10.1007/s10887-021-09199-6

Shen, Z., Zhang, K., & Dell, M. (2020). A Large Dataset of Historical Japanese Documents with Complex Layouts. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2336–2343. https://doi.org/10.1109/CVPRW50498.2020.00282

Sporleder, C. (2010). Natural Language Processing for Cultural Heritage Domains. *Language and Linguistics Compass*, *4*(9), 750–768. https://doi.org/10.1111/j.1749-818X.2010.00230.x

Stojkoski, V., Koch, P., & Hidalgo, C. A. (2023). Multidimensional economic complexity and inclusive green growth. *Communications Earth & Environment*, *4*(1), 130. https://doi.org/10.1038/s43247-023-00770-0

Stojkoski, V., Utkovski, Z., & Kocarev, L. (2016). The Impact of Services on Economic Complexity: Service Sophistication as Route for Economic Growth. *PLOS ONE*, *11*(8), e0161633. https://doi.org/10.1371/journal.pone.0161633

Turchin, P., Currie, T. E., Whitehouse, H., François, P., Feeney, K., Mullins, D., Hoyer, D., Collins, C., Grohmann, S., Savage, P., Mendel-Gleason, G., Turner, E., Dupeyron, A., Cioni, E., Reddish, J., Levine, J., Jordan, G., Brandl, E., Williams, A., … Spencer, C. (2018). Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. *Proceedings of the National Academy of Sciences*, *115*(2). https://doi.org/10.1073/pnas.1708800115

Van Der Wouden, F. (2020). A history of collaboration in US invention: Changing patterns of co-invention, complexity and geography. *Industrial and Corporate Change*, *29*(3), 599–619. https://doi.org/10.1093/icc/dtz058

Weber, I. M., Semieniuk, G., Westland, T., & Liang, J. (2021). What You Exported Matters: Persistence in Productive Capabilities across Two Eras of Globalization. *UMass Amherst Economics Department Working Paper Series*, *2021–02*. https://scholarworks.umass.edu/econ_workingpaper/299

# CHAPTER 2:
# THE ROLE OF IMMIGRANTS, EMIGRANTS, AND LOCALS IN THE HISTORICAL FORMATION OF EUROPEAN KNOWLEDGE AGGLOMERATIONS

# The Role of Immigrants, Emigrants, and Locals in the Historical Formation of European Knowledge Agglomerations*

*Philipp Koch[1,2], Viktor Stojkoski[1,3], César A. Hidalgo[1,4,5]*

[1] Center for Collective Learning, ANITI, TSE-R, IAST, IRIT, Université de Toulouse, 31000 Toulouse, France.
[2] EcoAustria – Institute for Economic Research, 1030 Vienna, Austria.
[3] Faculty of Economics, University Ss. Cyril and Methodius in Skopje, North Macedonia.
[4] Alliance Manchester Business School, University of Manchester, United Kingdom.
[5] School of Engineering and Applied Sciences, Harvard University, United States.

**Abstract**

Did migrants make Paris a Mecca for the arts and Vienna a beacon of classical music? Or was their rise a pure consequence of local actors? Here, we use data on more than 22,000 historical individuals born between the years 1000 and 2000 to estimate the contribution of famous immigrants, emigrants, and locals to the knowledge specializations of European regions. We find that the probability that a region develops or keeps specialization in an activity (based on the birth of famous physicists, painters, etc.) grows with both, the presence of immigrants with knowledge on that activity and immigrants with knowledge in related activities. In contrast, we do not find robust evidence that the presence of locals with related knowledge explains entries and/or exits. We address some endogeneity concerns using fixed-effects models considering any location-period-activity specific factors (e.g. the presence of a new university attracting scientists).

* This chapter is published in *Regional Studies* (doi: 10.1080/00343404.2023.2275571).

# Introduction

Migrants help carry knowledge across space[1–5], shaping the geography of cultural and economic activities[6–10]. But most studies documenting the role of migrants in the diffusion of knowledge use recent data on patents[6,11–19], research[13,20], or product exports[21], or analyze historical spillovers within activities[22–33], leaving questions about the role of migrants in the historical formation of knowledge agglomerations relatively unexplored.

To explore the role of migrants in the historical formation of knowledge agglomerations we use biographic data on more than 22,000 famous individuals—artists, physicists, explorers, philosophers, etc.—living in Europe between the years 1000 and 2000. We use this data to investigate how immigrants, emigrants, and locals explain the probability that famous individuals specialized in an activity—that was not yet present in a region—are born during the next century. That is, we study how the knowledge of migrants and locals contributes to explain, for example, Paris becoming the birthplace of painters and Vienna of composers.

We can explore these questions by creating measures of knowledge spillovers within and between locations and activities. Consider spillovers across locations within the same activity. The knowledge that migrants carry across borders may impact a location's ability to give birth to famous figures in the activity that the migrants specialize in. That is, immigrant mathematicians may increase the probability that a city or region begets famous mathematicians. Similarly, emigrating mathematicians may decrease that probability. To capture such spillovers, we identify whether a region experiences a larger than expected inflow or outflow of famous individuals specialized in an activity.

Now consider spillovers across both locations and activities. Migrants and locals specialized in an activity (e.g. mathematics) can impact a region's ability to give birth to famous figures in a related activity (e.g. physics). To capture such spillovers, we use measures of relatedness[34–38], which exploit information on the colocation of activities to estimate how "cognitively close" a location is to an activity.

During the past decades, measures of relatedness have been validated as robust predictors of the probability that countries, regions and cities enter or exit an activity, such as product exports[34,39,40], technologies[41–47], industries[48–51], and research areas[52–54]. Recent contributions to this literature have focused on unpacking relatedness by considering multiple channels[55–62]. For instance, does industry-specific or occupation-specific knowledge contribute to the growth and survival of firms?[58] Or do value chains or knowledge agglomerations explain the collocation of firms?[56] To the best of our knowledge no study has yet unpacked relatedness in the context of

historical migration. Here we use a dataset spanning 1,000 years of history in Europe to explore how the knowledge of immigrants, emigrants, and locals explains the probability that a famous cultural figure specialized in an activity is born in a specific region. This contributes to both, understanding the role of migrants in the geography of knowledge and unpacking relatedness metrics in the context of migration.

Our findings show that migrants play a crucial role in knowledge agglomerations. Specifically, we find that the probability that a European region enters a new activity grows on average by between 1.7 and 4.6 percentage points if that region received an excess number of immigrants specialized in that activity during the last century. Moreover, we find this correlation is enhanced by immigrants specialized in related activities. Similarly, we find the probability that a European region loses one of its existing specializations decreases on average by 5.0 to 10.2 percentage points if that region received an excess number of immigrants specialized in that activity. This correlation is also enhanced by immigrants specialized in related activities. In contrast, we do not find a statistically significant and robust role of the related knowledge of locals (people born in that region) in entries or exits.

To tackle some important endogeneity concerns (migration is often a motivated choice), we employ a highly restrictive fixed effects structure controlling for all possible unobserved factors that are specific to a broad occupational category in a region during a century. These are factors that might affect both, migration patterns and the birth of famous individuals, such as a new university attracting scientists and leading to the birth of more famous scientists in the future, or a prosperous city attracting and begetting more artists. In addition, we control for unobserved factors that are specific to a more granular occupational category in a century which might affect both migration and births (e.g. the emergence of a new technology (e.g. photography) begetting a new occupational category (photographers)). This captures, for instance, that musicians and singers are likely to have different migration and birth patterns across time than other artists such as painters or actors. Lastly, we tackle some concerns of reverse causality by focusing on excess migration and estimating the expected number of migrants in a location. Although we control for multiple possible observed and unobserved factors to limit endogeneity concerns, we want to stress that we are not able to make strictly causal claims.

Together, these findings advance our understanding of the role of immigrants, emigrants, and locals in the historical formation of knowledge agglomerations. They contribute to both, the literature on the role of migrants in knowledge diffusion[1–9,11–18,20–28] and the literature on relatedness[34–62]. Moreover, by developing measures of the related knowledge of migrants, we

combine both migration and relatedness in a framework that can be used to study how knowledge spillovers across space and across activities combine in more recent settings. Lastly, this study provides a long-term perspective on the evolution of regional specialisations in Europe, a perspective which is underrepresented in the field of economic geography[63].

# Data & Methods

## Data

We use the 2020 version of Pantheon[64], a publicly available dataset including information on famous individuals with a Wikipedia page in more than 15 different language editions. We focus on the 22,847 famous individuals born or died in Europe between the years 1000 and 2000. We choose Pantheon because it assigns individuals using a controlled taxonomy of 101 occupations, such as painter, writer, composer, physicist, chemist, mathematician, etc. Pantheon provides a good sectoral disaggregation compared to other datasets which either have few sectors[65] or use uncontrolled taxonomies with duplicate entries, e.g. film director and movie director[66]. This granularity is needed to construct measures of specialization and relatedness. The full taxonomy and descriptive statistics are provided in the Supplementary Materials (SM) section 1.1.

We use geographic coordinates to assign the place of birth and death of each biography to European administrative regions (NUTS-2 or regions of similar size for countries outside the EU, e.g. Russian Oblasts, see SM section 1.2). Figures 1a and 1b show the places of birth and death of all individuals in our dataset within the applied administrative borders. Due to a lack of data on the full trajectory of individuals, we follow the literature investigating migration patterns of famous individuals[65–68] and use places of birth and death as rough proxies for migration. Manual inspection of a random sample of 200 biographies revealed places of death to be a valid proxy of an important living place for around 90 percent of biographies and corresponded to a place of major impact for 75 percent of biographies (SM section 1.3).

Finally, we assign each individual to a century $t$ based solely on his or her year of birth. That is, a famous person who is born in the 18$^{th}$ century in Brussels and died in Paris (in the 18$^{th}$ or 19$^{th}$ century) is considered a local in Brussels and an immigrant in Paris in the 18$^{th}$ century. We choose this approach since we do not have information on the time of migration. We take this into account in the regression models by lagging the independent variables (see also SM section 1.1).

## Descriptive Statistics: Migration & Spatial Concentration Patterns

We find that most of the migration of famous Europeans over the past 1,000 years took place within countries and towards large cities (e.g. from smaller cities in France to Paris). Figures 1c and 1d visualize the migration network. Migration is common among famous individuals. In fact, going back to the 11$^{th}$ century, the share of migrants in our dataset never drops below 65 percent. In the 19$^{th}$ century, almost 80 percent of famous individuals in our dataset died in a different region than the one in which they were born (see Fig. 1e).

These migration patterns are not random but follow a process of preferential attachment, clustering individuals in major cities[65,68–72] and leading to a higher spatial concentration for places of death than birth. For instance, 416 famous individuals were born in Paris in the 19$^{th}$ century, but 934 died there (SM section 2.1).

We use information entropy $H$ to quantify the spatial concentration of births and deaths across regions. Information entropy (base 2) estimates the number of yes-or-no questions that we would need to answer—on average—to find the place of birth or death of an individual (see SM section 2.1). If deaths are more concentrated than births, we will need less questions to guess a place of death than one of birth. We can use entropy $H$ to estimate the effective number of places of birth or death as $E=2^H$, which is the number of regions effectively experiencing the birth or death of a famous individual.

Figure 1f shows the effective number of places of birth and death $E$ for each century. Prior to the 15$^{th}$ century, the spatial concentration of famous births and deaths was similar. But starting in the 15$^{th}$ century, places of death have become more spatially concentrated and places of birth more widespread. In fact, by the 19$^{th}$ century famous individuals were effectively born in more than 200 (out of 405) regions across Europe, while they effectively died in only 100 regions (Fig. 1f).

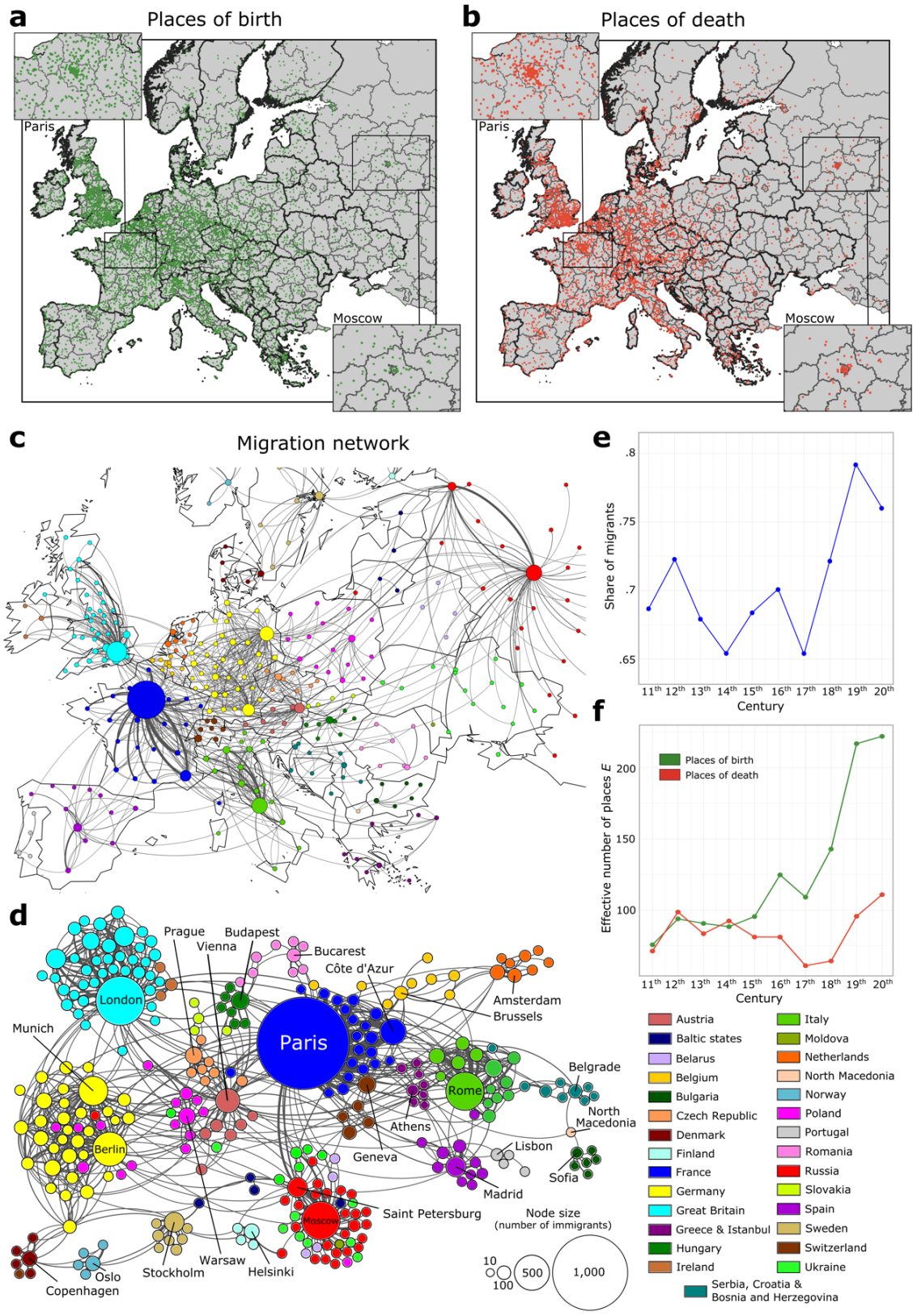



***Figure 1. Places of birth, places of death and migration patterns of famous individuals in Europe over the past 1,000 years.*** *(**a-b**) Maps of **a** places of birth and **b** places of death included in the analysis (NUTS-2 regions for EU, comparable regions for other countries, e.g. oblasts in Russia, see SM section 1.2). (**c-d**) Migration network of famous individuals within Europe over the past 1,000 years, using **c** geography or **d** a force-directed algorithm for visualization. The latter reveals that famous individuals tend to move within countries towards large regions. (**e**) Share of migrants in the dataset per century. (**f**) Effective number of places of birth and death E derived from Shannon entropy (see SM section 2.1). Starting in the 15th century, the places of death of famous individuals are more spatially concentrated than their places of birth.*

## Methods

### Relatedness of Immigrants, Emigrants, and Locals

To explore how the knowledge of immigrants, emigrants, and locals shapes the geography of knowledge, we estimate the probability that a region gives birth to a famous individual specialized in an activity as a function of estimates of knowledge spillovers within and between regions and activities.

To capture the knowledge spillovers of migration within the same activity, we calculate the ratio between the observed number of famous immigrants ($N_{ik,t}^{immi}$) or emigrants ($N_{ik,t}^{emi}$) with a certain activity and their expected number (respectively $\widehat{N}_{ik,t}^{immi}$ and $\widehat{N}_{ik,t}^{emi}$), where *i* denotes the region, *k* the occupation, and *t* the century.

Taking the ratio between the observed and expected number of migrants allows us to create measures of excess immigration or excess emigration, and thus, to control for the natural attractiveness of a location and the characteristics of an activity. This is important to address reverse causality concerns, since the effects of migrants could be simply a reflection of local factors making a place attractive for migrants with a certain specialization.

It is worth mentioning that migration decisions can be influenced by multiple local factors. Creatives, for instance, are more likely to move towards places that are already populated by other creatives[26] or potential patrons[73,74]. Geographical and cultural distance[75,76], such as a common language or the presence of fellow countrymen can also play a role[77]. Lastly, migration can also be exogenously forced due to conflict[28] or climate[78]. By focusing on excess migrants, instead of total migrants, in a restrictive fixed-effects model we help mitigate the risks of reverse causality.

Mathematically, this involves taking the ratio between the observed and expected number of immigrants or emigrants:

$$R_{ik,t}^{immi} = \frac{N_{ik,t}^{immi}}{\widehat{N}_{ik,t}^{immi}}$$

$$R_{ik,t}^{emi} = \frac{N_{ik,t}^{emi}}{\widehat{N}_{ik,t}^{emi}} \quad , \quad (1)$$

where the values are for individuals in region *i* and activity *k* born in century *t*.

Here we use two models for the expected number of migrants ($\widehat{N}_{ik,t}$). The first one considers the number of individuals in a location and the number of individuals specialized in an activity. That is a "bins and balls" model for the expected number of immigrants or emigrants, making

Eq. 1 the Revealed Comparative Advantage[79] or Location Quotient, a common measure of specialization:

$$\widehat{N}_{ik,t} = \frac{\sum_k N_{ik,t} \sum_i N_{ik,t}}{\sum_{i,k} N_{ik,t}}. \quad (2)$$

The second model expands on this by taking the attractiveness of a location in a specific activity into account[49]. We model $\widehat{N}_{ik,t}$ using a negative binomial regression where we control for the observed number in the previous century ($N_{ik,t-1}$), the previous specialization of the location in the activity based on famous individuals born there ($S_{ik,t-1}^{births}$),

$$S_{ik,t-1}^{births} = \frac{N_{ik,t-1}^{births}}{\left(\frac{\sum_k N_{ik,t-1}^{births} \sum_i N_{ik,t-1}^{births}}{\sum_{i,k} N_{ik,t-1}^{births}}\right)}. \quad (3)$$

where $N_{ik,t}^{births}$ denotes the number of famous individuals born in location *i* specialized in activity *k* in century *t*, and fixed effects for each location-time ($\theta_{it}$) and activity-time ($\vartheta_{kt}$) to account for unobserved factors. That is, we estimate

$$\widehat{N}_{ik,t} = f\left(\alpha_0 + \alpha_1 N_{ik,t-1} + \alpha_2 S_{ik,t-1}^{births} + \theta_{it} + \vartheta_{kt}\right), \quad (4)$$

where $f$ denotes the negative binomial probability density (see SM section 3.4.1 for results).

If the observed number of immigrants or emigrants in an activity exceeds the expected number, we say that region received excess immigrants, or produced excess emigrants, on that activity and time period.

Next, we create two specialization matrices for famous immigrants ($M_{ik,t}^{immi}$; died "here," but born elsewhere) and emigrants ($M_{ik,t}^{emi}$; born "here," but died elsewhere):

$$M_{ik,t}^{immi} = \begin{cases} 1 & if\ R_{ik,t}^{immi} \geq 1 \\ 0 & otherwise \end{cases}$$
$$M_{ik,t}^{emi} = \begin{cases} 1 & if\ R_{ik,t}^{emi} \geq 1 \\ 0 & otherwise. \end{cases} \quad (5)$$

Figures 2a and b show these two matrices using data for individuals born in the 19th century. The matrices are characterized by a nested structure that we recover by sorting locations by diversity (respectively $\sum_k M_{ik,t}^{immi}$ and $\sum_k M_{ik,t}^{emi}$), and activities by ubiquity (respectively $\sum_i M_{ik,t}^{immi}$ and $\sum_i M_{ik,t}^{emi}$). This structure is typical for matrices summarizing the geography of

activities[80,81] (SM section 2.2), but also, for networks describing species interactions in ecology[82–84].

To capture spillovers across activities we use measures of relatedness[34–38]. Relatedness exploits information on the colocation of activities to estimate their affinity with a location. We create three separate measures of relatedness for immigrants, emigrants, and locals.

These measures build on the specialization matrices described in Eq. 5. This time, however, we need to create specialization matrices for locals, which we define as famous individuals who were born in a region, no matter if they died there or elsewhere. We use this definition because of the large share of migrants among famous individuals (see Fig. 1d), which would reduce our number of observations drastically if we defined locals as individuals who were born and died in the same place. Controlling for the related knowledge of emigrants, however, relatedness based on all births is a valid proxy for the related knowledge of individuals who were born and died in the same region (SM section 2.3).

That is, as before, we calculate the ratio between observed and expected births of famous individuals:

$$R_{ik,t}^{births} = \frac{N_{ik,t}^{births}}{\widehat{N}_{ik,t}^{births}}. \tag{6}$$

Again, we can apply both the naïve model described in Eq. 2 or estimate the expected number of births given local factors (Eq. 3, see SM section 3.4.1) before creating binary specialization matrices for locals:

$$M_{ik,t}^{births} = \begin{cases} 1 & if\ R_{ik,t}^{births} \geq 1 \\ 0 & otherwise \end{cases} \tag{7}$$

This matrix also exhibits a nested structure (Fig. 2c).

Next, we define the proximity or similarity between two activities as the minimum of the conditional probability that a location is specialized in both of them[34]:

$$\varphi_{kk',t}^{immi} = \frac{\sum_i M_{ik,t}^{immi} M_{ik',t}^{immi}}{max\left(\sum_i M_{ik,t}^{immi}, \sum_i M_{ik',t}^{immi}\right)},$$

$$\varphi_{kk',t}^{emi} = \frac{\sum_i M_{ik,t}^{emi} M_{ik',t}^{emi}}{max\left(\sum_i M_{ik,t}^{emi}, \sum_i M_{ik',t}^{emi}\right)}, \tag{8}$$

$$\varphi_{kk\prime,t}^{births} = \frac{\sum_i M_{ik,t}^{births} M_{ik\prime,t}^{births}}{max\left(\sum_i M_{ik,t}^{births}, \sum_i M_{ik\prime,t}^{births}\right)},$$

and use these proximities to calculate the relatedness between locations and activities as:

$$\omega_{ik,t}^{immi} = \frac{\sum_{k'} M_{ik\prime,t}^{immi} \varphi_{kk\prime,t}^{immi}}{\sum_{k'} \varphi_{kk\prime,t}^{immi}},$$

$$\omega_{ik,t}^{emi} = \frac{\sum_{k'} M_{ik\prime,t}^{emi} \varphi_{kk\prime,t}^{emi}}{\sum_{k'} \varphi_{kk\prime,t}^{emi}}, \tag{9}$$

$$\omega_{ik,t}^{births} = \frac{\sum_{k'} M_{ik\prime,t}^{births} \varphi_{kk\prime,t}^{births}}{\sum_{k'} \varphi_{kk\prime,t}^{births}}.$$

These measures quantify how far, for example, immigrants to Paris are from being specialized in archeology, emigrants from Madrid are from being specialized in singing, or locals in Berlin are from being specialized in philosophy.

We note that the relatedness densities calculate with the naïve and binomial model are highly correlated ($R^2$>0.9). So, going forward, we present results using the naïve model and provide additional results using the negative binomial model in the SM (section 3.4.1).

Since the multiple factors contributing to the colocation of activities can be different when looking at immigration, emigration and births, we create separate measures of proximity ($\varphi_{kk\prime,t}^{immi}$, $\varphi_{kk\prime,t}^{emi}$, $\varphi_{kk\prime,t}^{births}$, Eq. 8). But as a robustness check, we also consider a joint measure of proximity ($\varphi_{kk\prime,t}^{joint}$) using colocation at birth and death (see SM section 2.5). Nevertheless, we find the separate measures of proximity provide valuable nuance (see SM Figure S6). Consider explorers and military personnel. Explorers and military personnel share many required capabilities such as navigating, planning, commanding etc., that may be explained by local factors such as military academies for education, distance to the sea, recency of a war, or naval technology. Also, exploration teams often involve soldiers and military personnel, which could then become famous as explorers. Hence, explorers and military personnel are likely to share a geographic origin. Yet, since exploration and military campaigns tend to involve different locations, these two activities are less likely to collocate at death. Now consider composers and noblemen. For these two activities, the proximity based on immigration patterns is higher than the proximity based on births. It makes sense that these activities are to some extent related when looking at places of birth: Noblemen are known to be patrons for the arts. Hence, noblemen born in a location will likely create institutions that promote the cultivation of the talent of composers born in this location. But it is also plausible that these activities are even

more related when looking at immigration patterns. Given that we observe a disproportional migration flow of noblemen towards a certain location, we can view this location as highly related to composers, since the institutional factors attracting noblemen likely play a role in attracting and cultivating the talent of composers as well.

These examples highlight why we believe that generating separate measures of proximity for immigrants, emigrants and births provides a nuanced perspective that helps unpack relatedness (see SM section 2.5. for more details).

We illustrate the structure of these proximity networks for immigrants born in the 19th century ($\varphi_{kk',t}^{immi}$, Fig. 2d). A high proximity between two activities indicates similarity or complementarity among them. Like measures of propensity, measures of proximity capture the combined presence of multiple factors that may be contributing to the colocation of two activities. For example, we find a high proximity between biologists and physicians, mathematicians and physicists, and musicians and actors (see Fig. 2d). While the latter may be considered an example of colocation due to high complementarity (musicians and actors may perform together), associations between mathematicians and physicists, or biologists and physicians, may indicate similarity in knowledge or skills.

**Entries and Exits**

We use our measures of relatedness to study the entry and exit of activities in European regions. We do this by estimating logistic models explaining the probability that a region starts to give birth to a disproportionately large number of famous individual specialized in an activity (entries) or stops doing so (exits). That is, a region enters the activity "philosophy" if more philosophers are born there in a certain century than expected, while this has not been the case in the prior century. Similarly, we explain the probability that a region loses an existing specialization. A region exits the activity "physics" if fewer physicists are born there than expected, while this has not been the case in the prior century. The variables $Entry_{ik,t}$ and $Exit_{ik,t}$ emerge directly from the specialization matrix defined in Eq. 7.

Specifically, we define:

$$Entry_{ik,t} = \begin{cases} 1 & if\ M_{ik,t-1}^{births} = 0\ and\ M_{ik,t}^{births} = 1 \\ 0 & otherwise \end{cases},$$
$$Exit_{ik,t} = \begin{cases} 1 & if\ M_{ik,t-1}^{births} = 1\ and\ M_{ik,t}^{births} = 0 \\ 0 & otherwise \end{cases}. \quad (10)$$

That is, a region *i* enters (exits) an occupation *k* in century *t* if the observed births of famous individuals with that occupation during the considered century is larger (lower) than expected, while this was not the case in the prior century.

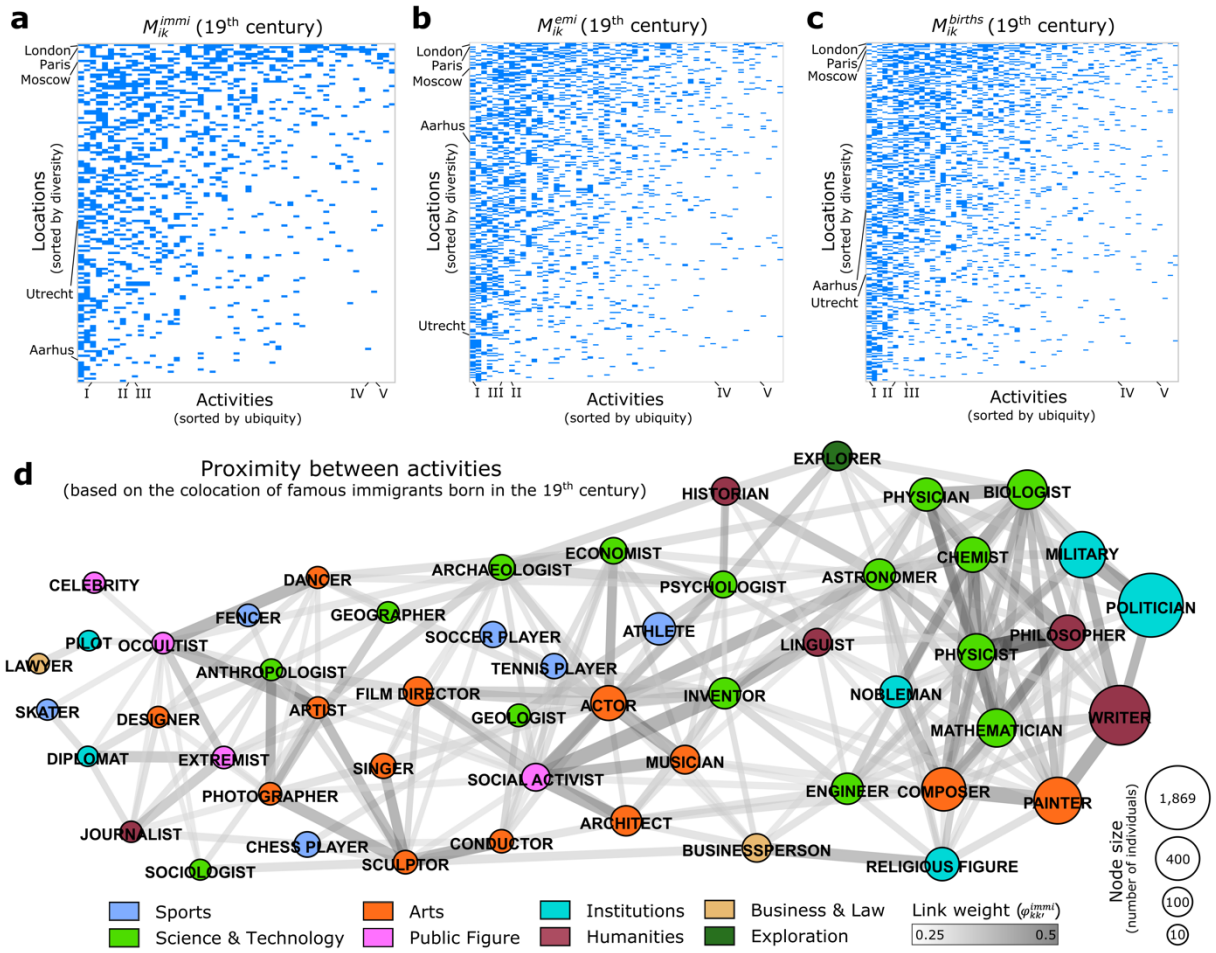



***Figure 2. Nested specialization matrices and the similarity between activities in the 19th century.*** *(**a-c**) Specialization matrices based on **a** immigrants, **b** emigrants, and **c** locals in the 19th century (see Eq. 5 and 7). Examples of locations and activities (I: writer, II: mathematician, III: physicist, IV: journalist, V: pilot) are highlighted. (**d**) Proximities between activities based on the colocation of famous immigrants born in the 19th century using the naïve model described in Eq. 2 to determine the expected number of immigrants. Node size is proportional to the number of famous individuals specialized in the respective activity and born in the 19th century.*

Defining entries and exits looking at places of birth is a rather conservative approach. For a region to enter an activity, it needs to become a place where the required knowledge to cultivate a certain talent can be absorbed through formal or informal institutions and social ties. Indeed, early exposure to local knowledge in an individual's life is highly relevant in shaping his or her career, both for inventors nowadays[85] and artists centuries ago[86]. A different approach of describing the geography of knowledge would be, for instance, to focus on all individuals living at a certain place. But this would require having data on all places of living.

We explain entries and exits using measures of the presence of immigrants and emigrants in that activity ($M_{ik,t}^{immi}$, $M_{ik,t}^{emi}$) and of the related activities that we can attribute to immigrants, emigrants and locals ($\omega_{ik,t}^{immi}$, $\omega_{ik,t}^{emi}$, $\omega_{ik,t}^{births}$). For instance, a significantly positive correlation between $M_{ik,t}^{immi}$ and entries would point towards migrants bringing into the region the knowledge needed to carry out activity *k*. That is, a high influx of mathematicians would increase the probability that the region begets its own famous mathematicians. This would be consistent with research showing that migrants help carry the knowledge needed to enter an activity[6,11–18,20–28].

Similarly, a significant correlation between $\omega_{ik,t}^{immi}$ and entries would support the idea that the related knowledge brought by migrants also impacts the probability that a region develops a new activity. That is, the knowledge of famous immigrants specialized in mathematics diffuses to related fields, such as physics or chemistry, and increases the probability that a region begets its own physicists or chemists.

Lastly, a significant correlation between $\omega_{ik,t}^{births}$ and entries, after controlling for $\omega_{ik,t}^{emi}$, would indicate that the related knowledge of locals contributes to entering a new activity. That is, a region with many locals already specialized in mathematics has a higher probability of branching into physics or chemistry.

The entry of a region into a new activity could be the result of multiple factors other than migration. For instance, the creation of a new university could attract scientists, and the expansion of a port could create conditions attractive to merchants. We address such endogeneity concerns by using highly restrictive fixed effects models accounting for unobserved factors that could affect both migration and the probability that a region enters an activity. Specifically, we control for these unobserved factors by using fixed effects specific to a broad occupational category, region, and century ($\gamma_{mit}$, i.e. a three-way interaction). Index *m* denotes one of eight broad occupational categories such as “arts”, “science & technology”, “humanities”, or “sports” (see column 1 of Table S1 in the SM).

In addition, we control for unobserved factors affecting both migration and future births that are specific to a more granular occupational category and time ($\delta_{lt}$). Index *l* denotes one of 26 occupation categories, which distinguish, for instance, between social sciences, natural sciences, and engineering within the broad category “science & technology” or music, design and film & theatre within the broad category “arts” (see column 2 in Table S1 in the SM). The latter fixed effects capture, for instance, that the invention of motion picture technology at the

end of the 19th century likely affected migration and birth patterns among film directors and actors differently than among other artists, such as painters or sculptors.

We also control for several other observed factors that might correlate with the probability of entry or exit and that are not captured in the fixed effects. This includes an activity's ubiquity (i.e. the number of locations specialized in it) and how close a region already is to having or losing a specialization ($R_{ik,t-1}^{births}$, see Eq. 6). Lastly, we account for knowledge diffusion across space due to other reasons than migration by creating measures of the spatial proximity to other regions with specializations in that specific activity or in related activities (see SM section 2.4). We provide descriptive statistics and discuss the explanatory variables in more detail in SM section 3.1.

In sum, we define $Y_{ik,t} = \{Entry_{ik,t}, Exit_{ik,t}\}$ and estimate:

$$P(Y_{ik,t}) = g(\beta_1 M_{ik,t-1}^{immi} + \beta_2 M_{ik,t-1}^{emi} + \beta_3 \omega_{ik,t-1}^{immi} + \beta_4 \omega_{ik,t-1}^{emi} + \beta_5 \omega_{ik,t-1}^{births} + \boldsymbol{\alpha}' \boldsymbol{X}_{ik,t-1} + \gamma_{mit} + \delta_{lt} + \varepsilon_{ik,t}) \quad , \quad (11)$$

where $g$ denotes the logistic probability density, $\boldsymbol{X}_{ik,t-1}$ denotes a vector of observed control variables and $\gamma_{mit}$, $\delta_{lt}$ the fixed effects.

We calculate average marginal effects based on this logistic regression by computing the marginal effect for each data point and taking the average.

# Results

Table 1 and Figure 3 show the relationship between the activities of immigrants, emigrants, and locals and the number of observed entries and exits. For entries (Table 1, columns 1-5), the probability correlates positively with an excess inflow of migrants specialized in an activity during the previous century ($M_{ik,t-1}^{immi} = 1$). Specifically, an excess of immigrants increases the probability of entry on average by 4.6 percentage points (Fig. 3a). Figure 3b plots the probability of entry as a function of $M_{ik,t-1}^{immi}$.

We also find that the probability of entry grows with the related knowledge of immigrants. A standard-deviation-increase of $\omega_{ik,t-1}^{immi}$ increases the probability of entry on average by 5.8 percentage points (Fig. 3a). Figure 3c visualizes the results by plotting the average probability of entry as a function of the relatedness density of immigrants ($\omega_{ik,t-1}^{immi}$). In accordance with the literature[35,87], the average probability of entry grows super-linearly, from 1.1 percent if no

related knowledge of famous immigrants is present in a region ($\omega_{ik,t-1}^{immi} = 0$) to 16.4 percent if all related activities are present ($\omega_{ik,t-1}^{immi} = 100$). Moreover, we find a positive correlation ($p < 0.1$) between $\omega_{ik,t-1}^{births}$ and entries, but unlike the estimate of the related knowledge of immigrants, this correlation is not robust (SM section 3.4).

When we look at exits (Table 1, columns 6-10), we find similar relationships but with the opposite sign. An excess inflow of famous individuals specialized in an activity during the previous century ($M_{ik,t-1}^{immi} = 1$) reduces the probability of exit significantly by 10.2 percentage points on average (Fig. 3d). Also, the related knowledge of immigrants ($\omega_{ik,t-1}^{immi}$) helps prevent losing specialization in an activity. Figure 3e and f visualize these results by plotting the probability of exit as a function of $M_{ik,t-1}^{immi}$ and $\omega_{ik,t-1}^{immi}$, respectively.

***Table 1***. *Main results of logistic regression models explaining entries and exits of activities.*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | Dependent Variable: $Exit_{ik,t}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| $M_{ik,t-1}^{immi}$ | 0.334*** | 0.303*** | 0.336*** | 0.331*** | 0.300*** | -0.603*** | -0.584*** | -0.591*** | -0.587*** | -0.571*** |
| | (0.080) | (0.075) | (0.086) | (0.080) | (0.076) | (0.127) | (0.134) | (0.120) | (0.126) | (0.126) |
| $M_{ik,t-1}^{emi}$ | 0.115 | 0.045 | 0.106 | 0.121 | 0.018 | 0.310 | 0.330 | 0.233 | 0.306 | 0.291 |
| | (0.261) | (0.278) | (0.261) | (0.255) | (0.270) | (0.240) | (0.232) | (0.216) | (0.222) | (0.203) |
| $\omega_{ik,t-1}^{immi}$ | | 0.027*** | | | 0.028*** | | -0.067*** | | | -0.064*** |
| | | (0.006) | | | (0.007) | | (0.016) | | | (0.011) |
| $\omega_{ik,t-1}^{emi}$ | | | -0.006 | | -0.024 | | | -0.048 | | -0.025 |
| | | | (0.012) | | (0.019) | | | (0.038) | | (0.063) |
| $\omega_{ik,t-1}^{births}$ | | | | 0.011 | 0.027* | | | | -0.059*** | -0.034 |
| | | | | (0.008) | (0.015) | | | | (0.018) | (0.041) |
| Further controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *Fixed effects:* | | | | | | | | | | |
| Broad categ.-region-century | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Category-century | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 3944 | 3944 | 3944 | 3944 | 3944 | 1051 | 1051 | 1051 | 1051 | 1051 |
| Pseudo-$R^2$ | 0.213 | 0.214 | 0.213 | 0.213 | 0.215 | 0.224 | 0.230 | 0.226 | 0.226 | 0.232 |
| BIC | 9537.0 | 9539.4 | 9545.0 | 9544.5 | 9553.1 | 3619.6 | 3618.0 | 3623.4 | 3623.3 | 3628.8 |

The fixed effects in these models are highly restrictive, amounting to more than 700 parameters in columns (1)-(5) and more than 350 parameters in columns (6)-(10). All regions included in the regression model exhibit a minimum number of births and migrants such that measures of specialization and relatedness are defined (see SM section 2.2). Standard errors are clustered by region and period. The full regression tables with all control variables are provided in SM Sections 3.2 and 3.3. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

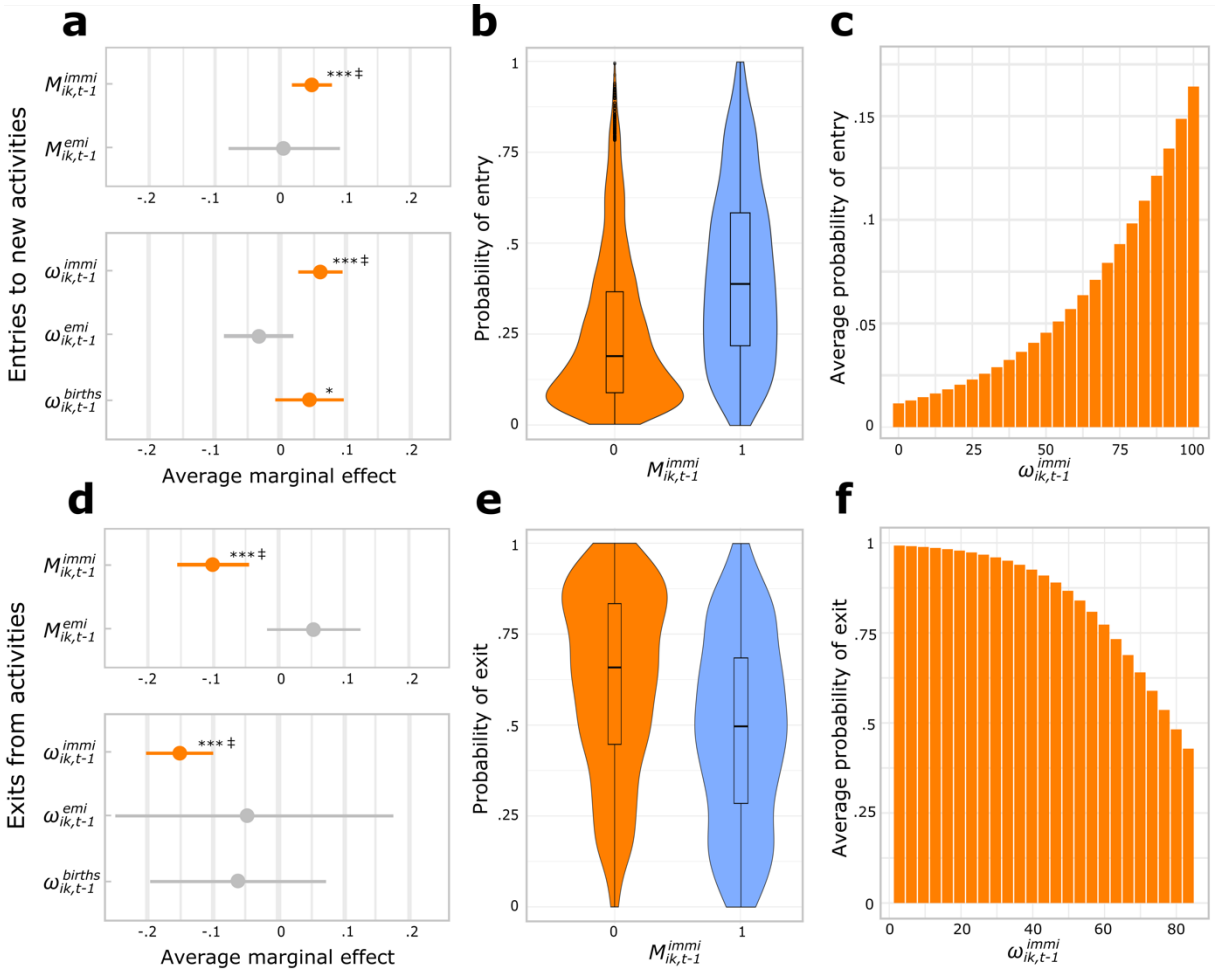



***Figure 3**. **Visualization of main results.** (**a**) Average marginal effects on the probability of entry to new activities based on the logistic regression model in Table 1, column 5. $M_{ik,t-1}^{immi} = 1$ increases the probability of entry on average by 4.6 percentage points, while a standard-deviation-increase of $\omega_{ik,t-1}^{immi}$ correlates with an increase in the average probability of entry of 5.8 percentage points. (**b-c**) Probability of entry to a new activity as a function of **b** $M_{ik,t-1}^{immi}$ and **c** the immigrants' related knowledge, $\omega_{ik,t-1}^{immi}$. (**d**) Average marginal effects on the probability of exit from activities based on the logistic regression model in Table 1, column 10. $M_{ik,t-1}^{immi} = 1$ reduces the probability of exit on average by 10.2 percentage points, while a standard-deviation-increase of $\omega_{ik,t-1}^{immi}$ correlates with a reduction in the average probability of exit of 15.1 percentage points. (**e-f**) Probability of exit from an existing area of specialization as a function of **e** $M_{ik,t-1}^{immi}$ and **f** the immigrants' related knowledge, $\omega_{ik,t-1}^{immi}$. Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$; Average marginal effects are computed by taking the average of the marginal effects across observations; Lines indicate 95% confidence interval; ‡ denotes robustness of the results (SM section 3.4)*

These results are robust to estimating the expected number of immigrants, emigrants, and locals in Eq. 1 and 6 using the negative binomial regression model described in Eq. 3 (see SM section 3.4.1). By accounting for local factors, we are able to obtain a more accurate estimate of the expected number of immigrants, emigrants, and locals and, thus, of a disproportionate migration flow. This mitigates some of the endogeneity concerns.

The highly restrictive fixed effects specification, however, reduces the number of observations in the regression model. To assure the robustness of our results, we estimate the logistic

regression models with several less restrictive specifications. This also allows us to include observed variables previously captured by the fixed effects, such as urban population[88,89] or a location's diversity of activities. We find that the knowledge of immigrants remains a significant predictor for both entries and exits (SM section 3.4.2). We acknowledge that, over such long periods, travel times are not constant but decrease with improvements in infrastructure and/or technology. Hence, we allow for century-specific effects of spatial proximity, $\rho_{ik,t-1}^{M}$ and $\rho_{ik,t-1}^{\omega}$, leaving our results unchanged (SM section 3.4.3). Also, our sample of famous individuals is not balanced over time. Our findings, however, are robust to excluding the 20th century from the analysis as well as looking at the 20th century alone (SM section 3.4.4). Moreover, our results do not change if we redefine entries and exits as the first or last birth of a famous individual with a specific occupation in a location instead of developing or losing specialization in an activity (SM section 3.4.5). In addition, we explore the explanatory power of interaction terms between various relatedness densities on entries, following the literature on migrants as agents of structural change[6–8]. We find a significant, but quantitatively negligible negative interaction term between $\omega_{ik}^{immi}$ and $\omega_{ik}^{births}$, indicating that the related knowledge of immigrants and locals are weak substitutes (SM section 3.4.6). Also, it may be that our findings of knowledge spillovers are different for different activities. We explore potential heterogenous effects by estimating our regression model separately for aggregate occupational categories (SM section 3.4.7). We find, for instance, a stronger correlation of the presence of immigrants specialized in the same activity ($M_{ik,t-1}^{immi}$) on entries in sciences and public institutions, and an increased correlation of related knowledge of immigrants ($\omega_{ik}^{immi}$) in humanities and sports. Another source of heterogeneity can be city size when size plays a relevant role in generating knowledge spillovers (SM section 3.4.8). We find that most entries take place in large cities, and thus, the effects of migration and relatedness are mainly urban. Whereas for exits, we see that spillovers across activities are more important in larger cities. But the probability of exiting an activity in small cities grows massively with the emigration of individuals specialized in the same activity, pointing towards a pronounced role of talent loss in shaping regional specialisations of small agglomerations. This finding relates to the recent literature on left-behind places[90,91].

Lastly, although the ratio of observed above expected numbers (Eq. 1 and 6) fulfils the purpose of controlling for size and reverse causality, these models are opaque, not telling us whether our results are driven by changes in the observed or expected number (or both). Hence, we run our main regression model including all terms of the ratio as a robustness check (SM section 3.4.9). We find that the observed number of immigrants with a specific occupation ($N_{ik,t-1}^{immi}$)

correlates positively with future entries and negatively with future exits, confirming our main results with composite indices. One additional immigrant to a region with a specific occupation correlates with an average increase in the probability of entry by 1.68 percentage points and a reduction in the probability of exit by 5.04 percentage points (SM section 3.4.9).

# Discussion

Labor mobility and migration are core tenets of the United States and the European Union, because policymakers intuit that migrants carry knowledge across space and activities[1–5]. Yet, despite multiple studies documenting the role of migrants in the diffusion of knowledge[6,11–33], there is little historical quantitative evidence of the role of migrants in the historical evolution of knowledge agglomerations.

Here, we used biographic data on more than 22,000 famous individuals—sculptors, composers, politicians, chemists, etc.—living in Europe between the years 1000 and 2000 to explore how the knowledge of immigrants, emigrants, and locals explains the probability that a region enters or exits an activity.

Our findings show that migrants play a crucial role in the historical geography of knowledge. Specifically, we find that the probability that a European region enters a new activity grows with the presence of immigrants with knowledge on that activity. Also, using measures of relatedness[34–38], we find that this correlation is enhanced by spillovers across related activities. Put differently, the probability that a region begets famous mathematicians grows with an excess immigration of mathematicians and with immigrants from related fields, such as physics or chemistry. Similarly, we find that the probability that a European region loses one of its existing areas of specialization decreases with the presence of immigrants specialized in that activity and in related activities. However, we do not find that locals with related knowledge play the same statistically significant and robust role in entries or exits.

These findings advance our understanding of the evolution of European agglomerations over the past millennium and of the role of migrants and locals therein. Specifically, we find robust evidence that European agglomerations did not only evolve path-dependently[92], but also that they benefited from spillovers generated by the migration of famous individuals. This supports the literature on the role of migrants in the diffusion of knowledge[1–8,11–18,20–28] and contributes to the literature on relatedness[34–38] explaining changes in specialization patterns[39–54].

Migrants are known agents of structural change enabling the development of unrelated activities[6–8]. Our findings differ slightly from that by emphasizing migration as a channel of

related diversification and path-dependent development, adding to the literature unpacking the principle of relatedness[55–62]. Recently, this intersection between evolutionary economic geography, regional diversification, and migration has been identified as a promising field of research[10]. We contribute methodologically to this literature by disentangling relatedness measures for immigrants, emigrants, and locals. These novel measures make it possible to explore how knowledge spillovers across space and across activities combine (SM section 2.5). Lastly, this study provides a long-term perspective on the evolution of regional specialisations in Europe, a perspective which has been underrepresented in economic geography[63].

Unfortunately, we do not observe the mechanisms explaining the entry or exit of regions in activities. There are, however, several potential mechanisms responsible for these results, which can be subsumed as horizontal and vertical socialization[93]. For instance, immigrating physicists could teach at a university, leading to a local flourishing of the field of physics and increasing the probability that a famous physicist emerges in the future. Also, immigrating physicists may bring new ideas and approaches with them, which can stimulate creative thinking and cross-pollination of ideas among local scientists in related fields such as chemistry or mathematics. This could lead to the development of new methods as well as new ways of thinking about problems, which could in turn contribute to an increased probability of giving birth to famous chemists or mathematicians in the future. The mechanisms may be different in other activities such as the arts or humanities. The presence of immigrating musicians may create a critical mass of artists, making it profitable to build cultural infrastructure due to economies of scale[26], from which artists in related activities such as singers, composers or dancers benefit as well. Shedding light on these different mechanisms is a promising avenue for future research.

Our study has also other limitations. First, we observe only a small and highly mobile subset of the overall population. That is, 22,000 of the most famous individuals living in Europe over the past 1,000 years. A more comprehensive dataset would allow for a more accurate and granular estimation of a location's related knowledge and the geography of activities. Indeed, we suspect that the limited sample is a likely reason for why we do not observe a statistically significant and robust relevance for locals in shaping the historical geography of knowledge. That being said, the related knowledge of locals plays a significant role in several specifications, for instance if estimating the expected number of famous individuals to define specialisations (SM section 3.4.1) or for large cities (SM section 3.4.8). Continuing to investigate the role of locals in the historical geography of knowledge can be an interesting avenue for future research.

Second, we do not observe the full migration trajectory of individuals, but only their place of birth and place of death. Although this approach follows the literature[65–68] and provides a good proxy of migration (SM section 1.3), more detailed data on where famous individuals lived and when could provide a better analytical basis to explore the evolution of agglomerations[94,95]. Indeed, based on a small number of famous individuals living between 1450 and 1750, it is estimated that they moved on average 3.72 times during their lifetime[96]. Third, we focus only on Europe. So, it may be that the principles behind the historical geography of knowledge uncovered here are different for other parts of the world. Lastly, migration is influenced by multiple factors such as geography and culture[75,76], agglomeration[26], patrons[73,74] or conflict[28,78], evoking reverse causality and endogeneity concerns in our study. We tackled these concerns by using highly restrictive fixed-effects and estimating the expected number of immigrants, emigrants, and locals to define specializations. Despite these efforts, we want to stress that we are not able to make strictly causal claims, a task that can be challenging using historical observational data.

Yet, despite these limitations, our study provides evidence of migration playing a central role in the evolution of European knowledge agglomerations. Also, while being a historical study, our study concerns a topic that is highly relevant in today's economic policy. The effects of migration on local economies have been debated intensively, both in academia[9,97–99] and in policy circles[100,101]. Our findings add to this debate by showing that the immigration of high-skilled individuals correlates with entering and exiting specialisations of regions. Yet, our results can neither be interpreted causally nor tell us whether these findings remain for migration that is incentivized by policy instruments, since we observe migration involving multiple forces, from forced displacement due to war, to organic forms of migration.

# Acknowledgements

We thank two anonymous referees, Andrea Morrison, Eva Coll, Jesús Crespo Cuaresma, Ron Boschma, Andrea Belmartino, the attendees of the 2022 Economic Geography PhD school in Utrecht, the attendees of the WICK#10 PhD Workshop in Economics of Innovation, Complexity and Knowledge in Turin, the attendees of the 13th Geoffrey J.D. Hewings Regional Economics Workshop in Vienna, members of the Complexity Science Hub and EcoAustria in Vienna, and the members of the Center for Collective Learning for valuable feedback.

## Funding

This project was supported by the Agence Nationale de la Recherche [grant number ANR-19-P3IA-0004], the 101086712-LearnData-HORIZON-WIDERA-2022-TALENTS-01 financed by the European Research Executive Agency (REA), and the European Lighthouse of AI for Sustainability [grant number 101120237-HORIZON-CL4-2022-HUMAN-02].


## Disclosure statement

The authors declare no competing interests.

## Data and materials availability

All data are available in the main text or the supplementary materials.

# References


1. Lissoni, F. International migration and innovation diffusion: an eclectic survey. *Reg. Stud.* **52**, 702–714 (2018).
2. Trippl, M. & Maier, G. Knowledge Spillover Agents and Regional Development. in *Innovation, Growth and Competitiveness* (eds. Nijkamp, P. & Siedschlag, I.) 91–111 (Springer Berlin Heidelberg, 2011). doi:10.1007/978-3-642-14965-8_5.
3. Williams, A. M. Lost in translation? International migration, learning and knowledge. *Prog. Hum. Geogr.* **30**, 588–607 (2006).
4. Kerr, S. P., Kerr, W., Özden, Ç. & Parsons, C. High-Skilled Migration and Agglomeration. *Annu. Rev. Econ.* **9**, 201–234 (2017).
5. Cipolla, C. M. The Diffusion of Innovations in Early Modern Europe. *Comp. Stud. Soc. Hist.* **14**, 46–52 (1972).
6. Miguelez, E. & Morrison, A. Migrant Inventors as Agents of Technological Change. *J. Technol. Transf.* (2022) doi:10.1007/s10961-022-09927-z.
7. Elekes, Z., Boschma, R. & Lengyel, B. Foreign-owned firms as agents of structural change in regions. *Reg. Stud.* **53**, 1603–1613 (2019).
8. Neffke, F., Hartog, M., Boschma, R. & Henning, M. Agents of Structural Change: The Role of Firms and Entrepreneurs in Regional Diversification. *Econ. Geogr.* **94**, 23–48 (2018).
9. Putterman, L. & Weil, D. N. Post-1500 Population Flows and The Long-Run Determinants of Economic Growth and Inequality. *Q. J. Econ.* **125**, 1627–1682 (2010).
10. Morrison, A. Towards an evolutionary economic geography research agenda to study migration and innovation. *Camb. J. Reg. Econ. Soc.* rsad013 (2023) doi:10.1093/cjres/rsad013.
11. Fassio, C., Montobbio, F. & Venturini, A. Skilled migration and innovation in European industries. *Res. Policy* **48**, 706–718 (2019).
12. Breschi, S., Lissoni, F. & Miguelez, E. Foreign-origin inventors in the USA: testing for diaspora and brain gain effects. *J. Econ. Geogr.* **17**, 1009–1038 (2017).
13. Bosetti, V., Cattaneo, C. & Verdolini, E. Migration of skilled workers and innovation: A European Perspective. *J. Int. Econ.* **96**, 311–322 (2015).
14. Hunt, J. & Gauthier-Loiselle, M. How Much Does Immigration Boost Innovation? *Am. Econ. J. Macroecon.* **2**, 31–56 (2010).
15. Miguélez, E. & Moreno, R. Research Networks and Inventors' Mobility as Drivers of Innovation: Evidence from Europe. *Reg. Stud.* **47**, 1668–1685 (2013).
16. Miguélez, E. & Moreno, R. Knowledge flows and the absorptive capacity of regions. *Res. Policy* **44**, 833–848 (2015).
17. Miguelez, E. & Noumedem Temgoua, C. Inventor migration and knowledge flows: A two-way communication channel? *Res. Policy* **49**, 103914 (2020).
18. Bahar, D., Choudhury, P. & Rapoport, H. Migrant inventors and the technological advantage of nations. *Res. Policy* **49**, 103947 (2020).
19. Bernstein, S., Diamond, R., Jiranaphawiboon, A., McQuade, T. & Pousada, B. The Contribution of High-Skilled Immigrants to Innovation in the United States. *NBER Work. Pap. Ser.* w30797 (2022) doi:10.3386/w30797.
20. Trippl, M. Scientific Mobility and Knowledge Transfer at the Interregional and Intraregional Level. *Reg. Stud.* **47**, 1653–1667 (2013).
21. Bahar, D. & Rapoport, H. Migration, Knowledge Diffusion and the Comparative Advantage of Nations. *Econ. J.* **128**, F273–F305 (2018).

22. Hornung, E. Immigration and the Diffusion of Technology: The Huguenot Diaspora in Prussia. *Am. Econ. Rev.* **104**, 84–122 (2014).
23. Moser, P., Voena, A. & Waldinger, F. German Jewish Émigrés and US Invention. *Am. Econ. Rev.* **104**, 3222–3255 (2014).
24. Ganguli, I. Immigration and Ideas: What Did Russian Scientists "Bring" to the United States? *J. Labor Econ.* **33**, S257–S288 (2015).
25. Diodato, D., Morrison, A. & Petralia, S. Migration and invention in the Age of Mass Migration. *J. Econ. Geogr.* **22**, 477–498 (2022).
26. Borowiecki, K. J. & Graddy, K. Immigrant artists: Enrichment or displacement? *J. Econ. Behav. Organ.* **191**, 785–797 (2021).
27. Mitchell, S. London calling? Agglomeration economies in literature since 1700. *J. Urban Econ.* **112**, 16–32 (2019).
28. Borowiecki, K. J. Are composers different? Historical evidence on conflict-induced migration (1816-1997). *Eur. Rev. Econ. Hist.* **16**, 270–291 (2012).
29. Waldinger, F. Quality Matters: The Expulsion of Professors and the Consequences for PhD Student Outcomes in Nazi Germany. *J. Polit. Econ.* **118**, 787–831 (2010).
30. Waldinger, F. Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany. *Rev. Econ. Stud.* **79**, 838–861 (2012).
31. Scoville, W. C. The Huguenots and the Diffusion of Technology. I. *J. Polit. Econ.* **60**, 294–311 (1952).
32. Scoville, W. C. The Huguenots and the Diffusion of Technology. II. *J. Polit. Econ.* **60**, 392–411 (1952).
33. Collins, H. M. The TEA Set: Tacit Knowledge and Scientific Networks. *Sci. Stud.* **4**, 165–185 (1974).
34. Hidalgo, C. A., Klinger, B., Barabási, A.-L. & Hausmann, R. The product space conditions the development of nations. *Science* **317**, 482–487 (2007).
35. Hidalgo, C. A. *et al.* The Principle of Relatedness. in *Unifying Themes in Complex Systems IX* (eds. Morales, A. J., Gershenson, C., Braha, D., Minai, A. A. & Bar-Yam, Y.) 451–457 (Springer International Publishing, 2018). doi:10.1007/978-3-319-96661-8_46.
36. Hidalgo, C. A. Economic complexity theory and applications. *Nat. Rev. Phys.* **3**, 92–113 (2021).
37. Boschma, R. Relatedness as driver of regional diversification: a research agenda. *Reg. Stud.* **51**, 351–364 (2017).
38. Balland, P.-A. *et al.* The new paradigm of economic complexity. *Res. Policy* **51**, 104450 (2022).
39. Poncet, S. & de Waldemar, F. S. Product Relatedness and Firm Exports in China. *World Bank Econ. Rev.* **29**, 579–605 (2015).
40. Pinheiro, F. L., Hartmann, D., Boschma, R. & Hidalgo, C. A. The time and frequency of unrelated diversification. *Res. Policy* 104323 (2021) doi:10.1016/j.respol.2021.104323.
41. Balland, P.-A., Boschma, R., Crespo, J. & Rigby, D. L. Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. *Reg. Stud.* **53**, 1252–1268 (2019).
42. Boschma, R., Balland, P.-A. & Kogler, D. F. Relatedness and technological change in cities: the rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Ind. Corp. Change* **24**, 223–250 (2015).
43. Rigby, D. L. Technological Relatedness and Knowledge Space: Entry and Exit of US Cities from Patent Classes. *Reg. Stud.* **49**, 1922–1937 (2015).

44. Petralia, S., Balland, P.-A. & Morrison, A. Climbing the ladder of technological development. *Res. Policy* **46**, 956–969 (2017).
45. Balland, P.-A. & Boschma, R. Complementary interregional linkages and Smart Specialisation: an empirical study on European regions. *Reg. Stud.* **55**, 1059–1070 (2021).
46. Juhász, S., Broekel, T. & Boschma, R. Explaining the dynamics of relatedness: The role of co-location and complexity. *Pap. Reg. Sci.* **100**, 3–21 (2021).
47. Uhlbach, W.-H., Balland, P.-A. & Scherngell, T. Public R&D funding and new regional specialisations: The contingent role of technological relatedness. *Ind. Innov.* **29**, 511–532 (2022).
48. Essletzbichler, J. Relatedness, Industrial Branching and Technological Cohesion in US Metropolitan Areas. *Reg. Stud.* **49**, 752–766 (2015).
49. Neffke, F., Henning, M. & Boschma, R. How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions. *Econ. Geogr.* **87**, 237–265 (2011).
50. Neffke, F., Otto, A. & Weyh, A. Inter-industry labor flows. *J. Econ. Behav. Organ.* **142**, 275–292 (2017).
51. Deegan, J., Broekel, T. & Fitjar, R. D. Searching through the Haystack:The Relatedness and Complexity of Priorities in Smart Specialization Strategies. *Econ. Geogr.* **97**, 497–520 (2021).
52. Boschma, R., Heimeriks, G. & Balland, P.-A. Scientific knowledge dynamics and relatedness in biotech cities. *Res. Policy* **43**, 107–114 (2014).
53. Guevara, M. R., Hartmann, D., Aristarán, M., Mendoza, M. & Hidalgo, C. A. The research space: using career paths to predict the evolution of the research output of individuals, institutions, and nations. *Scientometrics* **109**, 1695–1709 (2016).
54. Chinazzi, M., Gonçalves, B., Zhang, Q. & Vespignani, A. Mapping the physics research space: a machine learning approach. *EPJ Data Sci.* **8**, 33 (2019).
55. Farinha, T., Balland, P.-A., Morrison, A. & Boschma, R. What drives the geography of jobs in the US? Unpacking relatedness. *Ind. Innov.* **26**, 988–1022 (2019).
56. Diodato, D., Neffke, F. & O'Clery, N. Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time. *J. Urban Econ.* **106**, 1–26 (2018).
57. Bahar, D., Rosenow, S., Stein, E. & Wagner, R. Export take-offs and acceleration: Unpacking cross-sector linkages in the evolution of comparative advantage. *World Dev.* **117**, 48–60 (2019).
58. Jara-Figueroa, C., Jun, B., Glaeser, E. L. & Hidalgo, C. A. The role of industry-specific, occupation-specific, and location-specific knowledge in the growth and survival of new firms. *Proc. Natl. Acad. Sci.* **115**, 12646–12653 (2018).
59. Jun, B., Alshamsi, A., Gao, J. & Hidalgo, C. A. Bilateral relatedness: knowledge diffusion and the evolution of bilateral trade. *J. Evol. Econ.* **30**, 247–277 (2020).
60. Boschma, R. & Capone, G. Institutions and diversification: Related versus unrelated diversification in a varieties of capitalism framework. *Res. Policy* **44**, 1902–1914 (2015).
61. Zhu, S., He, C. & Zhou, Y. How to jump further and catch up? Path-breaking in an uneven industry space. *J. Econ. Geogr.* **17**, 521–545 (2017).
62. Cortinovis, N., Xiao, J., Boschma, R. & van Oort, F. G. Quality of government and social capital as drivers of regional diversification in Europe. *J. Econ. Geogr.* **17**, 1179–1208 (2017).
63. Henning, M. Time should tell (more): evolutionary economic geography and the challenge of history. *Reg. Stud.* **53**, 602–613 (2019).

64. Yu, A. Z., Ronen, S., Hu, K., Lu, T. & Hidalgo, C. A. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Sci. Data* **3**, 150075 (2016).
65. Schich, M. *et al.* A network framework of cultural history. *Science* **345**, 558–562 (2014).
66. Laouenan, M. *et al.* A cross-verified database of notable people, 3500BC-2018AD. *Sci. Data* **9**, 290 (2022).
67. De La Croix, D. & Licandro, O. The longevity of famous people from Hammurabi to Einstein. *J. Econ. Growth* **20**, 263–303 (2015).
68. Serafinelli, M. & Tabellini, G. Creativity over time and space: A historical analysis of European cities. *J. Econ. Growth* **27**, 1–43 (2022).
69. Borowiecki, K. J. Geographic clustering and productivity: An instrumental variable approach for classical composers. *J. Urban Econ.* **73**, 94–110 (2013).
70. O'Hagan, J. & Borowiecki, K. J. Birth Location, Migration, and Clustering of Important Composers: Historical Patterns. *Hist. Methods J. Quant. Interdiscip. Hist.* **43**, 81–90 (2010).
71. Borowiecki, K. J. & Dahl, C. M. What makes an artist? The evolution and clustering of creative activity in the US since 1850. *Reg. Sci. Urban Econ.* **86**, 103614 (2021).
72. O'Hagan, J. & Hellmanzik, C. Clustering and Migration of Important Visual Artists: Broad Historical Evidence. *Hist. Methods J. Quant. Interdiscip. Hist.* **41**, 121–136 (2008).
73. Haskell, F. *Maler und Auftraggeber: Kunst und Gesellschaft im italienischen Barock*. (DuMont, 1996).
74. *Die Kunst der Mächtigen und die Macht der Kunst: Untersuchungen zu Mäzenatentum und Kulturpatronage*. (Akademie Verlag, 2007).
75. Caragliu, A., Del Bo, C., de Groot, H. L. F. & Linders, G.-J. M. Cultural determinants of migration. *Ann. Reg. Sci.* **51**, 7–32 (2013).
76. Lewer, J. J. & Van Den Berg, H. A gravity model of immigration. *Econ. Lett.* **99**, 164–167 (2008).
77. Rephann, T. J. & Vencatasawmy, C. P. Determinants of the Spatial Mobility of Immigrants: Evidence from Sweden. *Rev. Reg. Stud.* **30**, 189–213 (2000).
78. Abel, G. J., Brottrager, M., Crespo Cuaresma, J. & Muttarak, R. Climate, conflict and forced migration. *Glob. Environ. Change* **54**, 239–249 (2019).
79. Balassa, B. Trade Liberalisation and Revealed Comparative Advantage. *Manch. Sch.* **33**, 99–123 (1965).
80. Bustos, S., Gomez, C., Hausmann, R. & Hidalgo, C. A. The Dynamics of Nestedness Predicts the Evolution of Industrial Ecosystems. *PLoS ONE* **7**, e49393 (2012).
81. Hausmann, R. & Hidalgo, C. A. The network structure of economic output. *J. Econ. Growth* **16**, 309–342 (2011).
82. Bascompte, J., Jordano, P., Melián, C. J. & Olesen, J. M. The nested assembly of plant–animal mutualistic networks. *Proc. Natl. Acad. Sci.* **100**, 9383–9387 (2003).
83. Almeida-Neto, M., Guimarães, P., Guimarães, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117**, 1227–1239 (2008).
84. Bastolla, U. *et al.* The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature* **458**, 1018–1020 (2009).
85. Bell, A., Chetty, R., Jaravel, X., Petkova, N. & Van Reenen, J. Who Becomes an Inventor in America? The Importance of Exposure to Innovation. *Q. J. Econ.* **134**, 647–713 (2019).
86. Galenson, D. W. *Conceptual revolutions in twentieth-century art*. (Cambridge University Press, 2009).

87. Alshamsi, A., Pinheiro, F. L. & Hidalgo, C. A. Optimal diversification strategies in the networks of related products and of related research areas. *Nat. Commun.* **9**, 1328 (2018).
88. Bairoch, P., Batou, J. & Chèvre, P. *La population des villes européennes de 800 à 1850*. (Librairie Droz, 1988).
89. Buringh, E. The Population of European Cities from 700 to 2000: Social and Economic History. *Res. Data J. Humanit. Soc. Sci.* **6**, 1–18 (2021).
90. Rodríguez-Pose, A. The revenge of the places that don't matter (and what to do about it). *Camb. J. Reg. Econ. Soc.* **11**, 189–209 (2018).
91. Rodríguez-Pose, A., Terrero-Dávila, J. & Lee, N. Left-behind versus unequal places: interpersonal inequality, economic decline and the rise of populism in the USA and Europe. *J. Econ. Geogr.* lbad005 (2023) doi:10.1093/jeg/lbad005.
92. Nunn, N. History as evolution. in *The Handbook of Historical Economics* 41–91 (Elsevier, 2021). doi:10.1016/B978-0-12-815874-6.00010-1.
93. Mokyr, J. *A Culture of Growth: The Origins of the Modern Economy*. (Princeton University Press, 2017). doi:10.1515/9781400882915.
94. Lucchini, L., Tonelli, S. & Lepri, B. Following the footsteps of giants: modeling the mobility of historically notable individuals using Wikipedia. *EPJ Data Sci.* **8**, 36 (2019).
95. Menini, S. *et al.* RAMBLE ON: Tracing Movements of Popular Historical Figures. in *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics* 77–80 (Association for Computational Linguistics, 2017).
96. Mokyr, J. Mobility, Creativity, and Technological Development: David Hume, Immanuel Kant and the Economic Development of Europe. in *Session on "Creativity and the Economy", German Association of Philosophy* (2005).
97. Borjas, G. J. The Economics of Immigration. *J. Econ. Lit.* **32**, 1667–1717 (1994).
98. Card, D. Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration. *J. Labor Econ.* **19**, 22–64 (2001).
99. Ottaviano, G. I. P. & Peri, G. Rethinking the Effect of Immigration on Wages. *J. Eur. Econ. Assoc.* **10**, 152–197 (2012).
100. OECD. *International Migration Outlook 2021*. (OECD, 2021). doi:10.1787/29f23e9d-en.
101. World Bank. *Leveraging Economic Migration for Development: A Briefing for the World Bank Board*. (World Bank, 2019).

# CHAPTER 3:
# AUGMENTING THE AVAILABILITY OF HISTORICAL GDP PER CAPITA ESTIMATES THROUGH MACHINE LEARNING

# Augmenting the availability of historical GDP per capita estimates through machine learning*

*Philipp Koch[1,2], Viktor Stojkoski[1,3], César A. Hidalgo[1,4,5]*

[1] Center for Collective Learning, ANITI, IRIT, Université de Toulouse, Toulouse, France.
[2] EcoAustria – Institute for Economic Research, Vienna, Austria.
[3] Faculty of Economics, University Ss. Cyril and Methodius, Skopje, North Macedonia.
[4] Center for Collective Learning, CIAS, Corvinus University, Budapest, Hungary.
[5] Toulouse School of Economics, Université de Toulouse, Toulouse, France.

## Abstract

Can we use data on the biographies of historical figures to estimate the GDP per capita of countries and regions? Here, we introduce a machine learning method to estimate the GDP per capita of dozens of countries and hundreds of regions in Europe and North America for the past seven centuries starting from data on the places of birth, death, and occupations of hundreds of thousands of historical figures. We build an elastic net regression model to perform feature selection and generate out-of-sample estimates that explain 90% of the variance in known historical income levels. We use this model to generate GDP per capita estimates for countries, regions, and time periods for which these data are not available and externally validate our estimates by comparing them with four proxies of economic output: urbanization rates in the past 500 years, body height in the 18th century, well-being in 1850, and church building activity in the 14th and 15th century. Additionally, we show our estimates reproduce the well-known reversal of fortune between southwestern and northwestern Europe between 1300 and 1800 and find this is largely driven by countries and regions engaged in Atlantic trade. These findings validate the use of fine-grained biographical data as a method to augment historical GDP per capita estimates. We publish our estimates with CI together with all collected source data in a comprehensive dataset.

## Significance statement

The scarcity of historical GDP per capita data limits our ability to explore questions of long-term economic development. Here, we introduce a machine learning method using detailed data on famous biographies to estimate the historical GDP per capita of hundreds of regions in Europe and North America. Our model generates accurate out-of-sample estimates ($R^2$ = 90%) that quadruple the availability of historical GDP per capita data and correlate positively with proxies of economic output such as urbanization, body height, well-being, and church building activity. We use these estimates to reproduce the reversal of fortunes experienced by southern and northern Europe and the historical role played by Atlantic ports. These findings show that machine learning can effectively augment the historical availability of economic data.

* This chapter is published in *PNAS* (doi: 10.1073/pnas.2402060121)

# Introduction

During the last decades, machine learning methods helped expand the economics toolbox (1, 2), from the use of satellite images to estimate poverty (3–6), population (7, 8), and land use (9–12), to the use of recommender systems to support economic diversification policies (13–16). But machine learning methods are not only useful to study the present or predict the future, they can also be used to explore the past. In this paper, we introduce a machine learning method designed to reconstruct historical GDP per capita estimates of dozens of European and North American countries and regions for the past 700 years, more than quadrupling the availability of historical economic output data for these regions.

For decades, economic historians have made great efforts to reconstruct the GDP per capita of countries and regions using historical documents on economic output (17, 18), and by approximating GDP per capita using data on consumption (19–26). Despite these efforts, estimates of historical GDPs per capita are still scarce (Figs. 1 A-B). The Maddison project, the largest collection of historical GDP per capita estimates (27, 28), has data for only 11 European countries for the year 1750 and 5 for the 1300s: France, England, Spain, Sweden, and Northern Italy. This leaves out important economies, such as those of Austria, Russia, and Switzerland in the 1750s, and those of most of Europe during the renaissance. GDP per capita estimates on a smaller geographic scale such as administrative regions or cities are even more scarce. For the year 1750, for instance, we only found regional GDPs per capita for Spain (29) and Sweden (30).

This lack of data limits our ability to explore questions of long-term economic growth and development. Yet, research on how to extend these estimates using big data and machine learning methods is still relatively unexplored. Here, we ask whether data on the biographies of hundreds of thousands of historical figures, combined with machine learning methods, can be used to extend GDP per capita estimates to countries, regions, and time periods for which this data is not available.

The use of data on historical figures is not a capricious choice. On the one hand, unlike data on GDPs per capita, there is an abundance of accurate biographical records. Recent research efforts have made available structured data on the places of birth, death, and occupations of hundreds of thousands of historical figures (31, 32), providing a potentially rich source of features that should correlate with regional variations in GDPs per capita. On the other hand, there are good reasons why the GDP per capita of a country or region should correlate with the probability that a historical figure is born or has died there.

Consider both direct and indirect channels. Inventors and scientists involved in productivity-enhancing and lifesaving innovations—such as James Watt and Alexander Fleming—may contribute directly to the GDP per capita of their economies (33–35) by increasing productivity or reducing disease burden. But there are also important indirect channels. Wealthier regions are more likely to attract talent, make local talent more visible, and provide the freedom and opportunities needed for individuals to specialize in cultural and economic activities. It is well known, for instance, that individuals that become famous—and get recorded historically—tend to be remarkably mobile (36–39). We should also expect these migratory forces to attract talented individuals to locations that are rich in terms of physical and human capital (39–48). For the sake of generating historical estimates of economic development, we are indifferent about whether wealth attracts talent, whether wealth makes talent more visible, or whether talent contributes directly to wealth. All of these channels imply a positive correlation with wealth that should be mineable from biographical records. In fact, our estimates do not require us to identify a causal link between any of these channels and GDP per capita, but to identify robust correlations between the presence of historical figures and the GDP per capita of the countries and regions where those individuals once located. That is, the careers of Michelangelo, Sandro Botticelli, and Filippo Lippi, tell us something about the prosperity of Tuscany in the 15$^{th}$ century, no matter whether they contributed to the wealth of Tuscany or were its by-products.

In this paper, we leverage information on more than 563K historical figures recorded across multiple languages in Wikipedia (31, 32) to test whether this data can be used to model the GDP per capita of hundreds of regions in Europe and North America for the past 700 years. Specifically, we train a set of supervised machine learning models (elastic net regression models) with geographical features derived from the biographies of famous historical figures to generate out-of-sample estimates of national and regional GDPs per capita (see Fig. 1 for a visual summary of the idea). We find the model provides encouraging results. In an out-of-sample test, it predicts the GDP per capita of European and North American countries and regions with an $R^2$=90.1% and a mean absolute error of 22.6% of the GDP per capita observed during that time period.

We externally validate these estimates by recreating qualitatively well-known historical development trajectories and by comparing them with other proxies of per capita wealth. First, we recreate the established finding that England and the Low Countries experienced larger economic growth than Southern Europe between 1300 and 1800 (49–53). We find that a large share of this reversal of fortune can be attributed to the rise of Atlantic trade, supporting earlier

findings by Acemoglu, Johnson & Robinson (49). Second, we show our estimates correlate with proxies of economic development, such as urbanization rates between 1500 and 1950 (54), body height in the 18th century (55), wellbeing in 1850 (56), and church building activity in the 14th and 15th century (57). These findings contribute a new method for the generation of historical GDP per capita estimates and open a door to the use of structured historical data for the estimation of long-term economic time series.

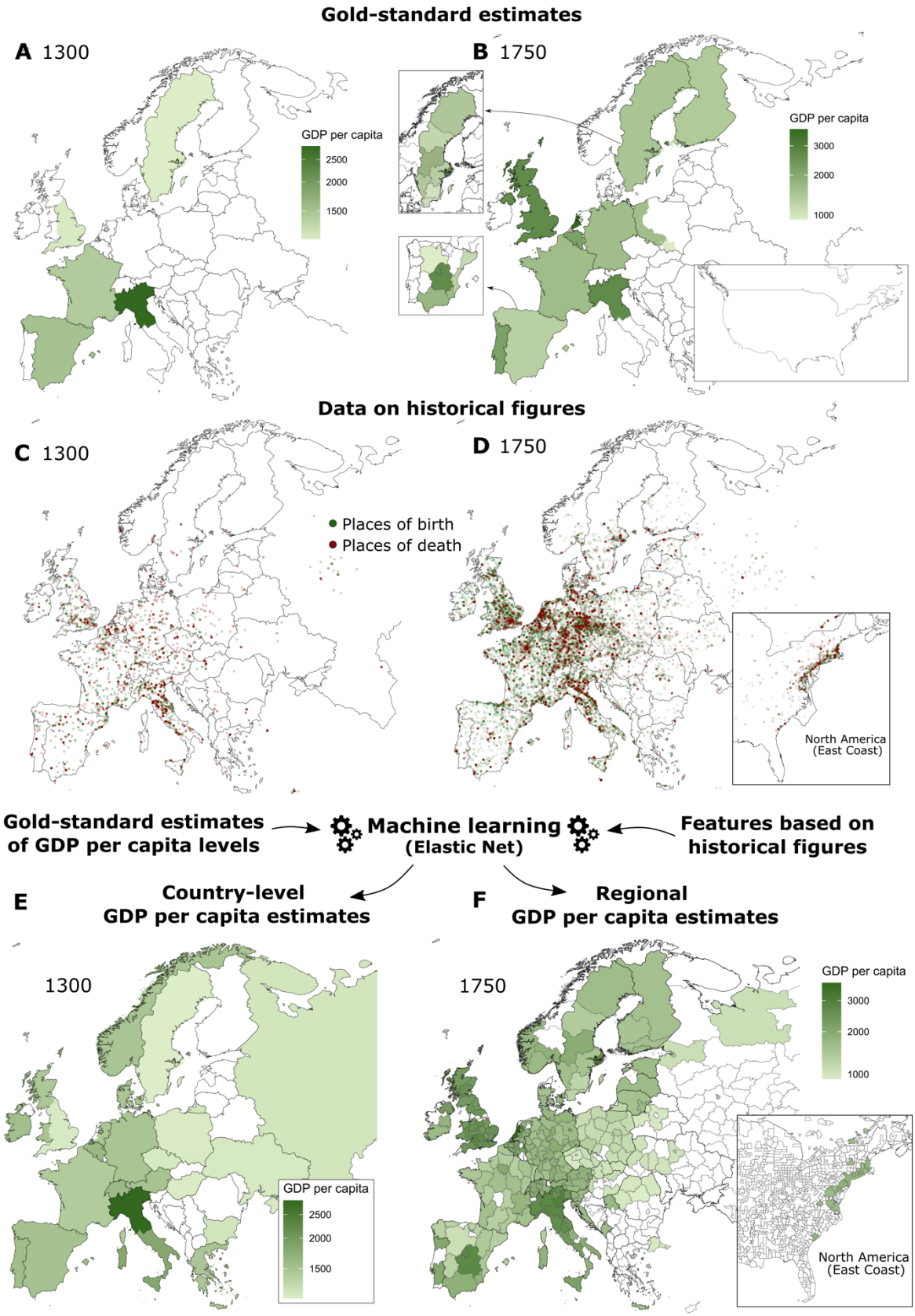



***Figure 1. Method Summary*** *(**A-B**) Gold-standard estimates on historical GDP per capita in (**A**) 1300 and (**B**) 1750 from the Maddison project and other sources for regional estimates (in 2011 USD). (**C-D**) Places of birth and death of famous individuals born at most 150 years prior to (**C**) 1300 and (**D**) 1750. (**E-F**) GDP per capita estimates based on available source data and machine learning models for (**E**) countries in 1300 and (**F**) regions in 1750.*

# Data

## Historical GDP per capita data

Our method builds on country-level GDP per capita estimates provided by the 2020 release of the Maddison project (27, 28). These are country-level estimates considering changing geographies. For instance, Great Britain data up to 1700 refers only to England (18), and data on Italy refers only to Northern Italy up to 1861 (20) (Figs. 1 A-B). For a full list of border changes, see the Maddison project (17–26, 29) and section 1 of the Supplementary Materials (SM).

We augment Maddison's country-level data with sources for estimates on the historical GDP per capita of regions (Fig. 1 B) in Spain between 1500 and 1800 (29), in Sweden between 1571 and 1950 (30, 58), in France in 1850 (59, 60), in the United Kingdom (61, 62) and Italy (63) between 1850 and 1950, and in Portugal (64) and Belgium (65) in 1900 and 1950.

Lastly, we add regional GDP per capita data for the year 2000 for most regions in the dataset. Specifically, we collect official data from Eurostat (66), the Office for National Statistics in the UK (67), the Bureau of Economic Analysis in the United States (68), Statistics Canada (69), the State Statistics Service of Ukraine (70), Belstat in Belarus (71), and Rosstat in Russia (72).

In total, we collect 1,336 GDP per capita observations in 50-year intervals (1300, 1350, …, 1950, 2000). All GDP per capita data is denoted in 2011 USD PPP, matching the unit provided in the Maddison project (SM section 1).

While the Maddison project is a comprehensive and widely used database on historical GDP per capita levels, its data must be understood as estimates. Comparing long-term economic development across the globe does not only require collecting and digitizing historical records, but also finding methods to compare purchasing powers across countries and continents. The latter are debated in the literature. For instance, it is argued that real income levels in the United States might have surpassed the ones in Europe earlier than data in the Maddison project claims (73, 74), or that the real income gap between Europe and Asia prior to the Industrial Revolution was far less pronounced (75). Despite this criticism, we use the Maddison project as gold-standard data since it has a large coverage and is still revised regularly by researchers at the University of Groningen (76).

## Data on historical figures

We use data on historical figures from a recently published database of notable people recorded on Wikipedia, curated and cross-verified by Laouenan et al. (31). This database contains

information on 2.29 million historical figures, including their places of birth, death, occupation, and proxies of their present-day popularity, such as Wikipedia page views or the number of language editions.

Data from Wikipedia is known to be subject to biases (77). For instance, famous figures of the Western world are overrepresented (78). Consider the 237K biographies Wikipedia provides of historical figures who are born between 1100 and 1900 (in at least two language editions and with an identifiable occupation). 77.9 percent of those biographies are about people who lived in Europe or North America. This is in contrast with population estimates showing that only 18.75 percent of the global population in 1820 lived in Europe or North America (27, 28). Also, cultural norms impact the portrayal of certain individuals in different language editions (79, 80), and the relative coverage of topics (81). Still, empirical studies find that the information available in Wikipedia is of relatively high accuracy when assessed by experts (82) or compared with other encyclopedias such as Britannica (83).

We address these limitations in two ways. First, we focus only on Europe and North America due to the limited representativity of other parts of the world. Second, we address potential language biases by considering only biographies with Wikipedia pages in at least two languages (to avoid including local biographies that are available only in a major language, such as English or French). We validate this methodological choice by comparing our results with models using data only from English pages or only non-English pages. We find similar results for all three samples suggesting that Wikipedia's English bias is not driving our estimates (SM section 5.5.2).

In total, we use 562,962 biographies of individuals living in Europe or North America after 1100 with an identifiable occupation and Wikipedia pages in at least two language editions (SM section 3.1). We assign biographies to countries and regions based on their places of birth and death (Fig. 1 C-D). To assign biographies to countries, we consider all border changes described in the source materials of the Maddison project (17–26, 29). For regions, we rely on European NUTS-2 regions (2021 edition), metro- and micropolitan statistical areas for the United States, metropolitan areas for Canada, and regions of similar size for other countries, e.g. oblasts in Russia (SM section 2.1). Finally, while the places of birth and death of historical figures do not provide a comprehensive view of their life history (e.g. Einstein was born in southern Germany and died in New Jersey, but lived also in Zurich and Berlin), they provide a proxy that has been used frequently in recent literature on historical migration (31, 36, 84, 85). In a recent publication (37), we tested this proxy by randomly sampling 200 individuals and manually

verifying their respective Wikipedia pages, finding that in 90% of the sample it was valid (SM section 3.4).

## Feature construction

We use this data to construct geographic features for each country, region, and time period. These include the total number of historical figures born in, died in, immigrated to (died in the place but born elsewhere), or emigrated from (born in the place but died elsewhere) each location; and occupation-specific counts (e.g. number of inventors or painters born, died, immigrated to, and emigrated from each location). These features are then weighted by an estimate of the historical popularity of each individual (the *Historical Popularity Index* (*HPI*) introduced in the Pantheon database (32)) and linearized using logarithms. *HPI* is an estimate of historical fame breaking the barriers of space, time, and language. It combines information on the number of Wikipedia pageviews, the number of language editions, and the age of historical figures (*Materials & Methods*). For a validation of the *HPI* see Yu et al. (32). Also, we calculate the average age of famous individuals, since increases in life expectancy have been shown to be leading indicators of the Industrial Revolution (85).

We augment this data with vectors generated using dimensionality reduction techniques such as singular value decomposition (SVD), a standard generalized eigenvalue decomposition for non-square matrices. We implement SVD by organizing our data into matrices describing the (*HPI*-weighted) number of historical figures in a location with a specific occupation. We create four different matrices for each year: births, deaths, immigrants, and emigrants, and include the first five eigenvectors of each matrix as candidate features. That is, we effectively include 20 SVD factors as potential candidate features for every year (*Materials & Methods*).

We also calculate estimates of economic complexity, an SVD type vector used frequently in economic development (16, 86, 87). The economic complexity index (*ECI*), is usually constructed with data on the geography of trade, employment, or patents, to explain cross-country and regional differences in economic growth (88–93), income inequality (92, 94) and greenhouse gas emissions (92, 95, 96). Here, we compute separate *ECI*'s for births, deaths, immigrants, and emigrants, and include them as potential features in our model (*Materials & Methods*). Finally, we include two more variables inspired by the literature on economic complexity: a location's diversity (the number of occupations with at least one individual in a location), and the average ubiquity of occupations in a location (the number of locations in which an activity, such as an occupation, is present).

Finally, there is the question of assigning features to time periods. For instance, which individuals should we consider when extracting features for the year 1600? In our model, we consider all individuals born in the 150 years prior to a respective year. That is, the features for 1600 include all biographies of individuals born between 1450 and 1600. We find our model is not too sensitive to this choice, as results using other thresholds (75, 100, and 175 years) are similar, but slightly worse than using the 150 years window (SM section 5.5.3).

In total, we collect between 250 and 300 potential features per period from the geography of famous biographies. In the next section we explain our feature selection process which is designed to avoid the risk of overfitting.

# Results

## Constructing the model

Armed with our data on historical figures and GDPs per capita, we now proceed to build and validate a model of GDP per capita estimates. To avoid overfitting, we use a regularized elastic net (EN) regression model (97). Elastic net models do not simply minimize the sum of squared residuals, like an OLS regression would, but penalize the model statistics using the $\ell^1$ and $\ell^2$ norms of the coefficients, effectively performing feature selection. This allows us to identify models that provide a good predictive power with an appropriate number of features.

We should note that the selected features can be different for different time periods. Attracting painters may be a positive predictor of GDP per capita in the 16$^{th}$ century but not in more recent years, and begetting inventors or engineers may be correlated with economic development during the Industrial Revolution but not during the renaissance. We take this into account by selecting features separately for each period. Since limited training data renders the selection for each year impossible, we perform feature selection for five historically informed time periods within which changes in importance are less likely. Specifically, we distinguish between the Late Middle Ages (1300-1500), the Early Modern Period (1501-1750), the Age of Revolutions (1751-1850), the Machine Age (1851-1950), and the Information Age (2000). Categorizing our analysis into these distinct periods allows us to capture changing relationships between the selected features and economic development.

For each period, we train the EN model with all available source data by optimizing the hyperparameters to find the most relevant features. We optimize the model's hyperparameters using k-fold cross validation and minimizing the prediction error (*Materials & Methods*). Then, we use this model to obtain out-of-sample estimates for countries and regions in Europe and

North America lacking GDP per capita data (Figs. 1 E-F). To avoid noise coming from the left-hand side of the distribution, we refrain from making predictions for locations with less than three births or deaths in a period up to 1600, with less than five births or deaths per period between 1650 and 1950, or with less than ten births or deaths in 2000. In total, we build upon our training data with 1,336 observations to provide out-of-sample estimates for 4,364 location-year combinations.

To make sure our regional estimates align with our country-level data, we rescale the regional estimates to match the population-weighted mean GDP per capita of the respective country. We use the number of births and deaths as population proxies, since data on historical population levels (54, 98) does not cover all regions in all periods and is restricted to urban population. The number of births and deaths is, however, a valid proxy of population (SM section 3.3). Lastly, we obtain standard errors and confidence intervals for our estimates by bootstrapping.

## Model performance

We assess model performance using out-of-sample cross-validation tests and by comparing it to a baseline model. For the out-of-sample cross-validation tests, we use withheld and independent test data sets. To ensure the test data sets are independent and minimize data leakage, we remove all observations for a randomly selected 20 percent of countries, including the regions within those countries (*Materials & Methods*).

Our baseline model accounts for persistence in income levels and differences between supranational regions (following the United Nations geoscheme, SM section 2.2). Specifically, it is a linear regression model that predicts GDP per capita with fixed effects for supranational regions in a specific period and the GDP per capita from the end of the previous historical period. The latter variable is not available for all locations and all time periods, so we use the following approach. If available, we use the GDP per capita at the end of the previous period from the source data. If that is not available, we use the estimates of the EN model in the previous historical period. For regions with unavailable source data or model estimates for the previous period, we use instead the data or estimates of the country that region is in. If none of the above is available, we use the average of the supranational region at the end of the previous period as initial GDP per capita. For example, the baseline prediction for the GDP per capita of Austria and Switzerland in 1800 accounts for the average GDP per capita of other Western European countries in 1800, as well as the GDP per capita of Austria and Switzerland in 1750.

The full model builds upon this baseline model and significantly improves it. Figs. 2 A and B are examples of how the fit improves compared to the baseline for one specific test data set

consisting of Italy, Portugal, Norway, Slovenia, Albania, Croatia, Romania and Latvia. For this test data set, the fit improves from 86% (baseline model) to 89% (full model). Figs. 2 C and D show the distribution of the R-squared and the mean absolute error across 500 different randomly selected independent test sets. The fit improves, at the median, from explaining 86.2% of the variance (baseline model) to 90.1% (full model), while the mean absolute error improves from 29% of average GDP per capita to 22.6%. Kruskal-Wallis H tests on statistical differences in the distributions between the baseline model and the full model are highly significant ($p < 1e^{-15}$). We provide further details on assessing model performance in the *Materials & Methods*.

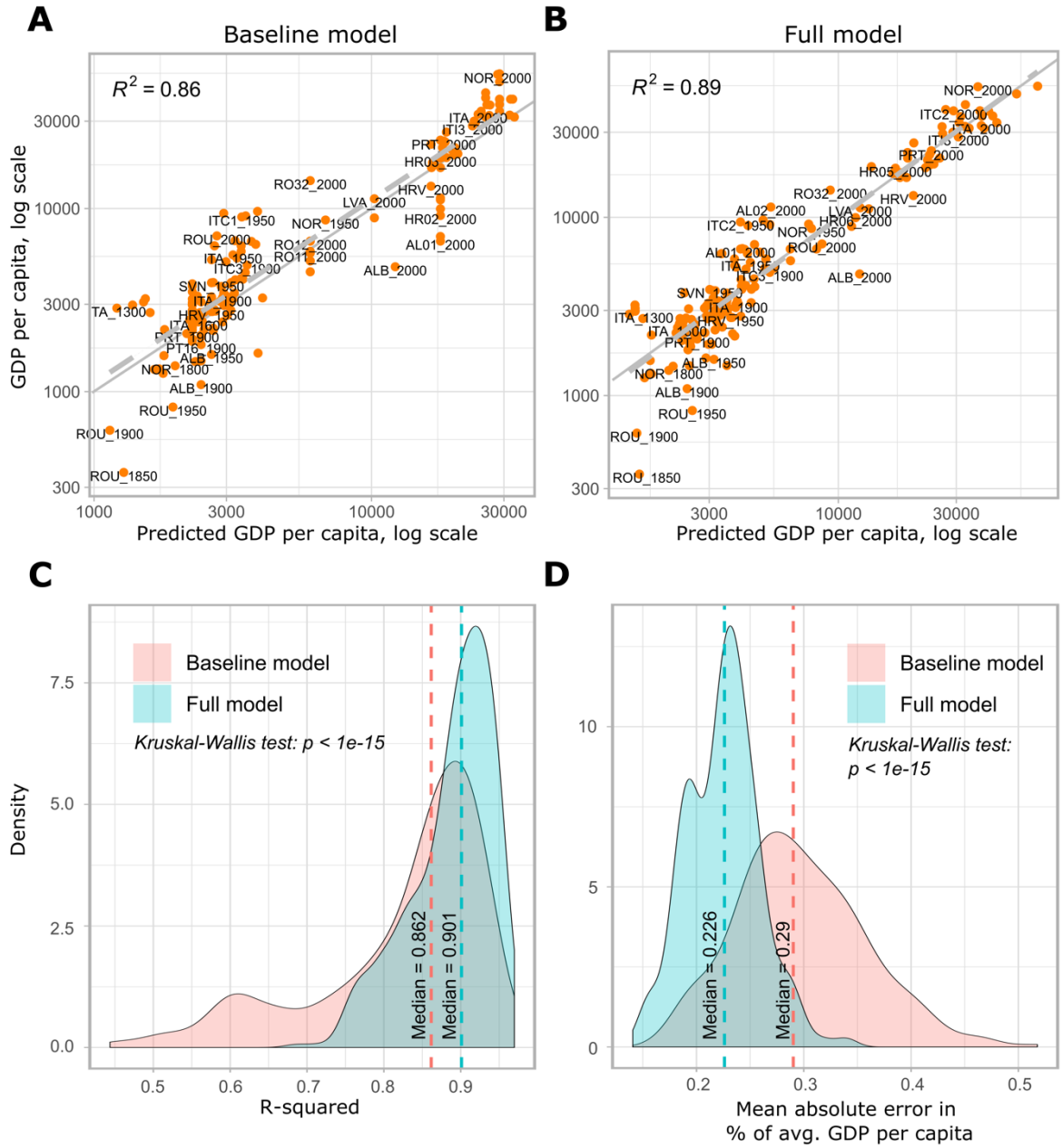



***Figure 2. Model performance.*** *(**A**) Baseline model prediction of test data for a random set of countries, accounting for fixed effects for supranational regions in a specific period (e.g. Southern Europe in 1950) and persistence in income levels. (**B**) Predictions of full model using elastic net. (**C**) Distribution of R-squared for the baseline and the full model when drawing 500 samples of training and test datasets. (**D**) Distribution of the mean absolute error for the baseline and the full model when drawing 500 samples of training and test datasets.*

## External validation: little divergence, urbanization, body height, wellbeing, and church building

We externally validate our estimates in two ways: First, we recreate Europe's well-known Little Divergence (49–53) and explore the role Atlantic trade therein (49). The Little Divergence refers to the observation that England, Netherlands, and Belgium experienced faster economic growth than Southern European countries (Italy, Spain, and Portugal) during the centuries leading to the Industrial Revolution. A central explanation for this divergence is the rise of

Atlantic trade starting the 16th century. Atlantic trade led to larger direct economic gains and shifted political power towards commercial interests. As Acemoglu et al. argue (49), the latter was not the case in countries with strong absolutist powers, which is why Spain and Portugal profited less from Atlantic trade than England and the Netherlands.

Our regional GDP per capita estimates reproduce these observations (Fig. 3 A-D). While Lombardy was one of the richest regions in Europe up to 1500, with an estimated GDP per capita of around 3,000 2011$, Amsterdam and London experienced higher economic growth in the following centuries. In 1800 Amsterdam and London were among the richest regions in Europe (Fig. 3 A).

To investigate the within-country variation of the Little Divergence we generate population-weighted deciles of GDP per capita for the North (England, Netherlands, Belgium) and the South (Italy, Spain, Portugal). We use the number of births and deaths of famous individuals in a location as population proxies (SM section 3.3). Our estimates show that the North experienced sustained economic growth between 1300 and 1800, while the South stagnated. Also, we find that, in 1300, the bottom 10th percentile of the South has been as rich as the top 90th percentile of the North. In 1800, the opposite holds: The bottom 10th percentile of the North exhibits a similar income level as the 90th percentile of the South (Fig. 3 B).

We show that Atlantic ports were a significant driver of this development. In line with results Acemoglu et al. (49), we find that countries with Atlantic ports (UK, NLD, FRA, ESP, PRT) experienced more rapid growth between 1300 and 1850 than other European countries (Fig. 3 C). Moreover, we find that this development is to a large extent driven by regions with Atlantic ports in the United Kingdom and the Netherlands (Fig. 3 D). Their average GDP per capita increased fivefold between 1300 and 1850, from 1,200 to 6,000 USD. In contrast, regions with Atlantic ports in France, Portugal, and Spain, and regions with Mediterranean ports did not experience sustained economic growth during the same period. This supports Acemoglu et al.'s (49) findings using city population as a proxy for regional economic development (SM section 5.2).

Second, we externally validate our estimates by showing they correlate with four known proxies of economic development: (a) urbanization rates between 1500 and 1950 (54), (b) average body height in the early and late 18th century (55), (c) a composite indicator of well-being in 1850 published by the OECD (56), and (d) city-level church building activity in cubic meters between 1300 and 1450 in Italy, France, Switzerland, the Low Countries, and Great Britain (57). We measure urbanization as the share of urban population (54) relative to total population according

to the Maddison project (27, 28). Indeed, urbanization is a frequently employed proxy of pre-industrial living standards and prosperity (49, 99, 100), as is body height (101–103). The OECD well-being indicator aggregates information on GDP per capita, wages, life expectancy, income inequality, years of education, homicide rates, and body height (56). And church building activity is associated with income levels because such projects have been major long-term investments, requiring a positive outlook on the future and the technological advances necessary for such endeavors. In all four cases we find our estimates correlate with these proxy measures (Fig. 3 E-H). We also find these correlations are very similar for labeled and unlabeled observations, alleviating some of the concerns with respect to the generalizability of our results (SM section 5.3).

Additionally, we explore whether our estimates can recreate patterns of regional development in German regions after the French Revolution as described by Acemoglu and coauthors (104). They find that German regions occupied by the French revolutionary armies, who induced radical institutional changes, experienced larger economic growth (proxied using urbanization rates) in the second half of the 19$^{th}$ century than other German regions. We replicate their descriptive plots with our regional estimates of GDP per capita and find highly similar patterns (SM section 5.4).

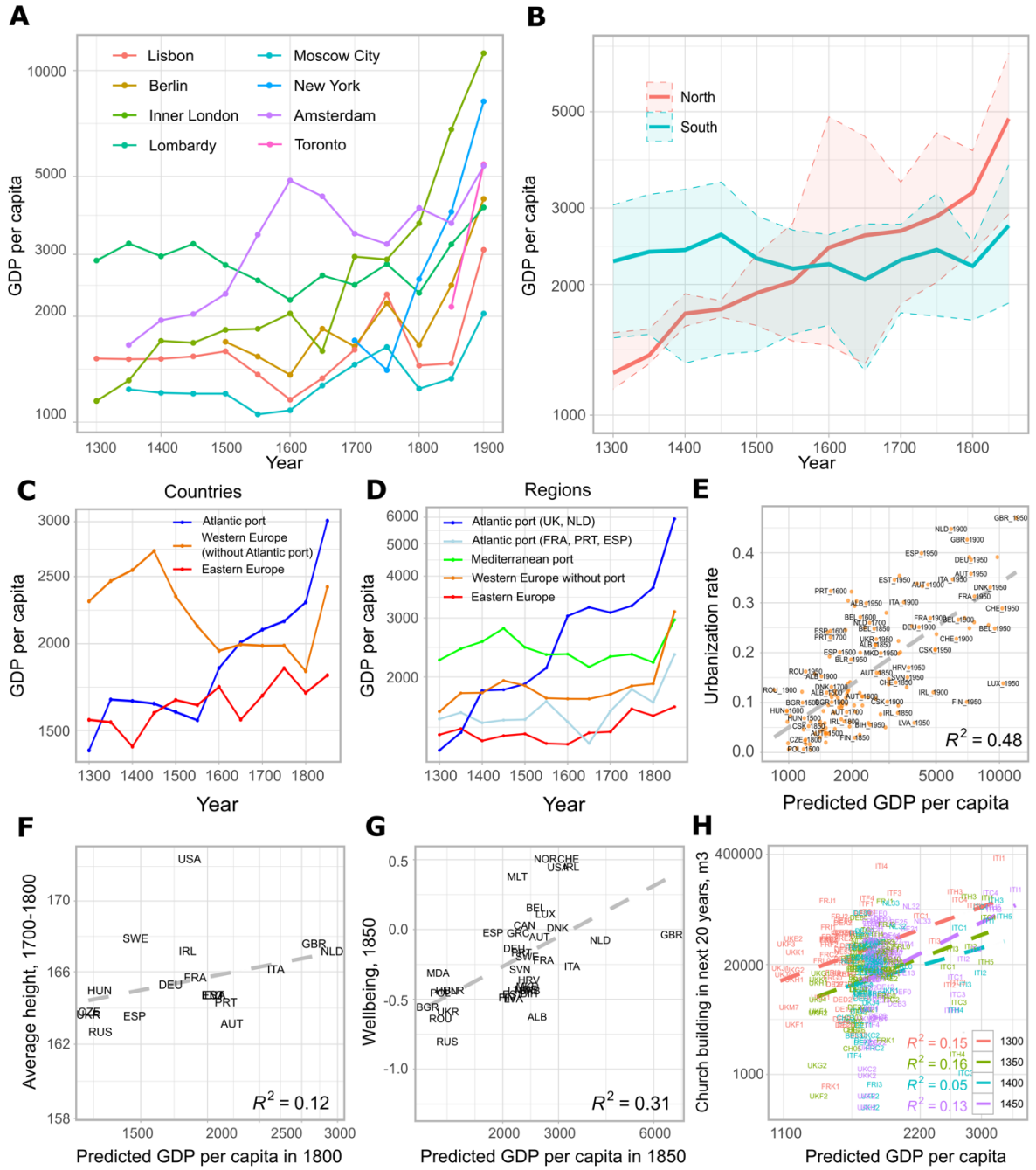



***Figure 3. External model validation.*** *(**A**) Economic development in selected European and North American regions and cities between 1300 and 1900. (**B**) Little Divergence: England, Netherlands, and Belgium (North) experienced sustained economic growth prior to the Industrial Revolution, while Italy, Spain, and Portugal (South) did not. Displayed are the population-weighted 90$^{th}$ and 10$^{th}$ percentiles, and the mean of the respective GDP per capita. (**C**) Economic development in countries with Atlantic ports, other Western European countries, and Eastern European countries (**D**) Economic development in regions with Atlantic ports, Mediterranean ports, and without a port, showing that Atlantic trade is a relevant driver of the Little Divergence. (**E**) Correlation of predicted GDP per capita with urbanization rates between 1500 and 1950. (**F**) Correlation of predicted GDP per capita with average body height in the 18$^{th}$ century. (**G**) Correlation of predicted GDP per capita with an indicator of wellbeing in 1850 published by the OECD. (**H**) Correlation of predicted GDP per capita with city-level church building activity in the 14$^{th}$ and 15$^{th}$ century.*

## Unpacking the evolution of prosperity in Europe and North America

Having validated our estimates, we now use them to explore some additional stylized facts. On the level of countries, our dataset provides several GDP per capita time series which were yet unavailable, such as Portugal prior to 1530, South Italy prior to 1861, Switzerland prior to 1850, Russia prior to 1885, Austria prior to 1820, and many more. Also, our estimates differentiate the British Isles countries prior to 1700, showing that England was the richest among them after 1400.

Figs. 4 A-C show the evolution of country-level economic development in Europe and North America between 1300 and 1900. In 1300, income levels have been highest in Northern Italy (Fig. 4 A). While the Netherlands and Belgium were among the richest economies in 1600 (Fig. 4 B), we find the United Kingdom and the United States to exhibit the highest income levels in 1900 (Fig. 4 C).

We now move from national to regional estimates of GDP per capita, which are even scarcer in published resources. Our dataset enables the investigation of economic development in Europe and North America on a regional level (Figs. 4 D-F). The overall findings are in line with the country-level estimates: Northern Italy became gradually less rich relative to other economies, while the Low Countries and the UK grew sharply. Regional estimates, however, provide more nuance. While the GDP per capita level in Spain was similar to France or England in 1600, we estimate income levels for Madrid (~2,600 USD) to be significantly higher than in London (~2,000 USD) or Paris (~1,800 USD), and even slightly higher than in regions in Northern Italy. Also, we find income levels in Amsterdam in 1600 (~4,900 USD) to be more than 30 percent above other parts of the Netherlands such as Rotterdam (~3,500 USD) or Utrecht (~2,200 USD). In 1900, income levels are more similar across Europe, with Great Britain topping the European charts. The richest cities back then, however, are found in the United States: According to our estimates San Jose and Los Angeles (~13,000 USD) had higher income levels in 1900 than Inner London (~11,200 USD).

When exploring our estimates, we found three insights that showcase potential use cases of our data.

First, we know from the Maddison project that Germany was one of the richest economies in Europe in 1500, prior to the Protestant Reformation. But which were the richest regions in Germany back then? Our estimates show that Nuremberg was the region within Germany with the highest GDP per capita in 1500 (Fig. 4 G). The city's prominent position is in line with historical research describing Nuremberg in the 16$^{th}$ century as a renaissance city and cultural

and economic center (104, 105). German income levels then fell between 1500 and 1600 on average by 29.6 percent. Nuremberg experienced a similar economic decline, according to our estimates. In contrast, regions that were relatively rich in 1500 but did not experience such a significant decline in the 16th century are Swabia (with its capital Augsburg) and Rheinhessen-Pfalz (incl. the cities Frankenthal and Kaiserslautern). One possible explanation is that cities in Swabia and Rheinhessen-Pfalz adopted Protestantism relatively early in the Reformation, and Protestantism has (ever since Max Weber) been connected to positive economic outcomes (106, 107). The link between Protestantism and economic prosperity, however, is not unquestioned. A empirical analysis of 272 cities in the Holy Roman Empire shows that there is no association between Protestantism and population growth (108). Here, we find that Protestant regions such as Nuremberg, Swabia and Rheinhessen-Pfalz were among the richest regions in 1500 and the latter two experienced less economic decline in terms of income per capita over the course of the 16th century than other German regions.

Second, we can use our estimates to explore the history of Charleston, South Carolina. Charleston emerged as a commercial hub and major city between 1720 and 1730 (109). We find it to be one of the richest metropolitan areas in North America in 1750 (Fig. 4 H). After the American Revolution, Charleston was the largest city in the South, continuing to be a center for slave trade (110). Our estimates reflect that since Charleston did not develop as positively as other cities during the antebellum era (Fig. 4 H).

Third, we find that the GDP per capita of Lisbon declined sharply after 1750 (Fig. 3 A). This observation coincides with the disastrous earthquake that hit Portugal's South in 1755 and had severe economic consequences (111). The Maddison project estimates the GDP per capita of Portugal fell by 33.2% between 1750 and 1800, and we estimate Lisbon's GDP per capita fell by 37.2% in this period. In contrast, we find that the GDP per capita of regions in Portugal that were not as affected by the earthquake even developed positively: Income per capita grew by 6.6% in Northern Portugal and by 9.5% in the region Alentejo.

## Feature importance

Finally, we explore the importance of the features selected by our model before providing additional evidence about the robustness of our results. We unpack feature importance using Shapley values. Shapley values originate from game theory (112) and are frequently applied in machine learning to interpret predictions (113, 114). These are defined as the average marginal effect of including a certain feature over all possible feature combinations (*Materials & Methods*).

Figs. 4 I-K show the most relevant features in 1300, 1600, and 1900, respectively. In 1300, the dummy variable for Eastern Europe is the most relevant feature, correlating negatively with GDP per capita. Looking at interpretable features derived from biographies, we find that being a place of deaths for famous lawyers and painters, and a place of birth for famous politicians are among the most relevant positive predictors of GDP per capita in 1300 (Fig. 4 I). In 1600, we find that the GDP per capita in the previous period is the most relevant feature in predicting GDP per capita levels. Also, the number of deceased and immigrant priests correlates negatively with income levels, while the number of deceased, born, and immigrant painters correlates positively (Fig. 4 J) with GDP per capita. We also find some SVD factors to be relevant features in 1600, such as the third factor describing the geography of famous births and the fourth factor describing the geography of famous deaths (SM section 4.2). These abstract factors, however, lack a direct interpretability compared to the number of births and deaths of individuals with a given occupation. In 1900, next to the initial income level, the diversity of occupations as well as the average age of famous individuals in a location are positive predictors of income levels (Fig. 4 K).


| Region (NUTS code) | GDP p.c. in 1500 | Δ GDP p.c., 1500-1600 |
|---|---|---|
| Nuremberg (DE25) | $ 2.049 | -29,8% |
| Freiburg (DE13) | $ 1.976 | -34,5% |
| Thüringen (DEG0) | $ 1.968 | -53,0% |
| Karlsruhe (DE12) | $ 1.965 | -36,4% |
| Leipzig (DED5) | $ 1.913 | -36,8% |
| Brandenburg (DE40) | $ 1.908 | -46,0% |
| Gießen (DE72) | $ 1.903 | -33,3% |
| Dresden (DED2) | $ 1.899 | -47,5% |
| Schleswig-Holstein (DEF0) | $ 1.833 | -39,7% |
| Rheinhessen-Pfalz (DEB3) | $ 1.829 | -25,1% |
| Swabia (DE27) | $ 1.827 | -14,8% |
| **Germany** | **$ 1.827** | **-29,6%** |

***Figure 4. Evolution of prosperity in Europe and North America.*** *(**A-C**) Country-level GDP per capita estimates in Europe and North America in (**A**) 1300, (**B**) 1600, and (**C**) 1900. (**D-F**) GDP per capita in European and North America regions and cities in (**D**) 1300, (**E**) 1600, and (**F**) 1900. (**G**) Richest regions in Germany in 1500 and economic growth in the 16th century. (**H**) Economic development of selected metropolitan areas in the United States between 1650 and 1900. (**I-K**) Feature importance measured in Shapley values for (**I**) 1300, (**J**) 1600, and (**K**) 1900.*

## Robustness of our estimates

Before we conclude, we check the sensitivity of our results to biases in the data and justify our methodological choices through several robustness checks (SM section 5.5). First, we investigate how our model performs when we use only data prior to the year 2000, since relatively recent time periods may upward bias our model's performance measures. We find this is not the case. While the $R^2$ is lower, model performance in terms of the mean absolute error improves slightly when we remove data for the year 2000 (SM section 5.5.1). Second, we investigate whether the English bias in Wikipedia significantly affects our estimates. For this purpose, we compare our results to those obtained using only English Wikipedia pages or only non-English Wikipedia pages. All three samples yield highly similar results, even for regions in English-speaking countries, indicating that this data limitation is not driving our estimates (SM section 5.5.2). Third, we provide model performance results for other thresholds of assigning biographies to time periods. We find that other thresholds yield similar, but slightly worse results than using 150 years (SM section 5.5.3). Fourth, we linearize our features before fitting our regression models. We use logarithms in our main results but provide robustness checks using the inverse hyperbolic sine function. We find that both scaling functions yield similar results (SM section 5.5.4). Fifth, we test whether backward feature selection performs better than EN regression models. We find that backward feature selection performs significantly worse (SM section 5.5.5). Sixth, we test whether our model is sensitive to the use us of HPI when deriving features from the biographies of historical figures, and find that removing the HPI slightly decreases model performance (SM section 5.5.6). Seventh, we test to what extent the dummies for supranational regions are driving our results. Removing them from the elastic net model yields only slightly worse results (SM section 5.5.7). Lastly, we investigate whether we can predict GDP per capita growth rates instead of levels following the same methodology. Here, we do not find a significant improvement compared to the baseline, a fact that could come from the significantly lower number of observations we have for growth (we need two observations for each growth number, meaning that we have only 455 ground truth observations for growth compared to over 1,300 for income levels) (SM section 5.5.8).

# Discussion

Despite significant efforts to collect data on historical income levels (27–29, 58–65), our understanding of long-term economic development remains limited. Here, we explored whether data on the biographies of historical figures can be used to create models of historical GDP per capita levels for countries and regions in Europe and North America for the past 700 years and

estimate their historical GDP per capita. Leveraging information on more than 563k historical figures recorded across multiple languages in Wikipedia (31, 32) we were able to construct a supervised machine learning model that makes relatively accurate predictions. In an out-of-sample test, this model predicts the GDP per capita of European and North American countries and regions with an $R^2$=90.1% and a mean absolute error of 22.6% of the GDP per capita observed during that time period.

We externally validated our estimates by recreating the finding of the Little Divergence (49–53), emphasizing the role of Atlantic trade in European economic development between 1300 and 1850, and showing that our estimates correlate with other proxies of economic development. The Little Divergence describes the fact that England and the Low Countries experienced larger GDP per capita growth rates in the centuries prior to the Industrial Revolution than Southern Europe. Our estimates confirm this finding and provide additional insights by enabling the comparison of within-country income distributions. We find that the bottom 10$^{th}$ percentile of the South has been as rich as the top 90$^{th}$ percentile of the North in 1300, while the opposite holds in 1800. Additionally, we find this is particularly driven by British and Dutch regions with Atlantic ports, supporting previous findings in the literature (49). Also, we find that our estimates correlate with four proxies of economic development: urbanization rates between 1500 and 1950 (54), body height in the 18$^{th}$ century (55), wellbeing indicators in 1850 (56), and city-level church building activity in the 14$^{th}$ and 15$^{th}$ century (57).

Armed with these estimates, we explored some stylized facts about economic development in Europe and North America that go beyond existing country-level estimates. For instance, we find income levels in Madrid in 1600 to be significantly higher than in Paris or London, despite an overall similar GDP per capita in Spain, France, and England. Moving to 1900, we find San Jose and Los Angeles had higher GDPs per capita than any other city in our dataset. We also explored the history of Nuremberg and other Protestant cities in the 16$^{th}$ century (104, 105), the development of Charleston, SC, in the 19$^{th}$ century (109, 110), and the economic consequences of the disastrous Lisbon earthquake in 1755 (111) as potential use cases of our estimates.

This method is, however, not without limitations. First, our data on GDP per capita levels going back centuries must be understood as estimates of estimates. That is, the "ground-truth" data we use to generate out-of-sample estimates are already estimates. This induces a level of uncertainty that needs to be considered when using our data and method. Second, data from Wikipedia is known to be subject to biases (77). We provide several robustness checks to show that our estimates are not affected by these biases and are careful to not extend our estimates to

Africa, Asia or South America. Third, we provide results using elastic net models since they are efficient in selecting features and preventing overfitting, but future research may come up with better models and methods. Lastly, countries and regions for which source data is available are not perfectly representative of locations without available source data. Indeed, countries and regions with source data tend to have a higher GDP per capita in 2000 and a higher number of famous individuals than countries and regions without source data. Still, we find that the correlation between our estimates and proxies of economic development is comparable for labeled and unlabeled observations, which alleviates some of the concerns with respect to the generalizability of our results (SM section 5.3).

Together, this paper introduces a new method for the generation of historical GDP per capita estimates with encouraging results and showcases the use of structured historical data for the estimation of long-term economic time series. Specifically, our findings validate the use of fine-grained biographical data as a method to produce historical GDP per capita estimates. We hope future research can build upon this idea to further improve our understanding of economic development. We publish our estimates with confidence intervals together with all collected source data and the code to replicate our results. This dataset does not only allow for investigating 700 years of cross-country differences in economic development, but also for comparing the development of different regions in Europe (Milan, Montpellier, Paris, London, etc.) with metropolitan areas in North America (New York, Boston, Toronto etc.).

# Materials & Methods

**Historical Popularity Index**. We take the historical popularity of individuals in our dataset into account when defining features. We reconstruct a version of the Historical Popularity Index (*HPI*) introduced in the Pantheon database (32) with available data. Specifically, an individual's *HPI* is proportional to the number of Wikipedia page views ($V$), the number of language editions ($L$) and age ($A$, i.e. 2023 minus year of birth):

$$HPI = \begin{cases} log_{10}(V) + ln(L) + log_4(A) & if\ A \geq 70 \\ log_{10}(V) + ln(L) + log_4(A) - \dfrac{70 - A}{7} & if\ A < 70 \end{cases}$$

This measure of historical popularity is strongly correlated with the *HPI* in the Pantheon dataset (which also includes information on the entropy of the distribution of pageviews across languages and uses information on pageviews in non-English editions of Wikipedia) ($R^2 = 0.76$, SM section 3.2).

**Economic Complexity**. To calculate economic complexity, we create binary adjacency matrices $M_{ik,t}$ which indicate whether a location is specialized in an occupation based on measures known as the Revealed Comparative Advantage or Location Quotient:

$$M_{ik,t} = \begin{cases} 1 & if\ \frac{N_{ik,t}/N_{i,t}}{N_{k,t}/N_t} \geq 1 \\ 0 & otherwise \end{cases},$$

where $N_{ik,t}$ denotes the number of famous individuals in location *i* with occupation *k*, weighted by their $HPI$. Then, the economic complexity index ($ECI$) is defined as the result of an iterative mapping, defining a location's complexity as the average complexity of the occupations it is specialized in:

$$ECI_i = \frac{1}{M_i} \sum\nolimits_k M_{ik} PCI_k$$

$$PCI_k = \frac{1}{M_k} \sum\nolimits_i M_{ik} ECI_i \quad .$$

We compute separate $ECI$s for births, deaths, immigrants, and emigrants (SM section 4.1).

**Singular Value Decomposition**. Singular Value Decomposition (SVD) is a dimensionality reduction technique which retrieves factors from a rectangular matrix that best explain its structure. Here, we collect our data in adjacency matrices $N^j_{ik,t}$ describing the (HPI-weighted) number of births, deaths, immigrants, or emigrants in a certain location with a certain

occupation. Index *i* denotes the location, *k* denotes the occupation and *j* differentiates between births, deaths, immigrants, and emigrants.

Mathematically, SVD decomposes matrix *N* (technically, we use its logarithm) into

$$N = U \times S \times V^T \text{ ,}$$

where $U$ and $V^T$ are unitary matrices collecting orthonormalized eigenvectors describing locations and occupations, respectively, and *S* is a diagonal matrix collecting the singular values (115). We include the first five eigenvectors in $U$ for births, deaths, immigrants, and emigrants as candidate features, i.e. twenty potential features per period (SM section 4.2).

**Elastic Net**. We use elastic net (EN) regression models to perform feature selection and generate out-of-sample estimates. EN does not simply minimize the sum of squared residuals like an OLS regression would do, but also penalizes for the $\ell^1$ and $\ell^2$ norms of the coefficients, effectively performing feature selection and reducing the risk of overfitting. Mathematically, the EN estimator $\widehat{\boldsymbol{\beta}}$ minimizes the following function $L$ for given parameters $\alpha$ and $\lambda$:

$$L(\alpha, \lambda, \boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2) \text{ ,}$$

where $\boldsymbol{y}$ is the log of GDP per capita (base 10) and $\boldsymbol{X}$ is a vector of features. Note that the EN collapses to a LASSO (least absolute shrinkage and selection operator) if $\alpha = 0$ and to a ridge regression if $\alpha = 1$. The parameter $\lambda$ controls the extent of the penalty. We find optimal values for $\alpha$ and $\lambda$ using k-fold cross-validation ($k = 10$), minimizing the prediction error. Parameter values and selected features for each period are provided in SM section 5.1.

**Model performance.** We test how well our model performs on out-of-sample data using 500 randomly drawn, independent test sets. Specifically, one iteration (out of 500) of assessing the model's performance consists of, first, randomly selecting 20% of countries. For the model performance to be accurate and unbiased, it is crucial to make sure the test set (the withheld 20% of countries) is independent of the choice of hyperparameters. Hence, we now use the remaining 80% of countries to tune the hyperparameters of the EN model (α and λ).

For tuning α and λ, we use 10-fold cross-validation. That is, the sample of 80% of countries is split into 10 subsamples. Then, we find hyperparameters by, iteratively, leaving one of those subsamples out (validation sets) and using the remaining 9 subsamples as training sets. The optimal hyperparameters are the averages over these 10 iterations. Next, we use this model (trained on 80% of the countries) to predict the GDP per capita of the remaining 20% of countries, which the model has not encountered yet, and compare our estimates with the respective source data (using R-squared and mean absolute error). We compute the R-squared

using the estimates for the log of GDP per capita, and use the exponentiated estimates for computing the mean absolute error. This procedure is repeated 500 times, eventually yielding Fig. 2 C-D.

**Shapley values**. Shapley value $\phi_i$ is defined as the average marginal effect of including feature $i$ in the model for all possible feature combinations $S$:

$$\phi_i = \sum_{S \subseteq F_{-i}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_{S \cup i}(x_{S \cup i}) - f_S(x_S)\,] \quad ,$$

where $F$ denotes the set of all model features.

**Data and code availability.** We publish our out-of-sample estimates together with the collected source data on countries (27, 28) and regions (29, 58–65) in a comprehensive dataset comprising 5,700 observations (1,336 source data observations, and 4,364 out-of-sample estimates). For the out-of-sample estimates, we provide 90 percent confidence intervals. Also, we publish the code to ensure reproducibility of our results. Data and code are available at https://github.com/philmkoch/historicalGDPpc.

# Acknowledgements


## Funding

We acknowledge the support of the Agence Nationale de la Recherche grant number ANR-19-P3IA-0004, the 101086712-LearnData-HORIZON-WIDERA-2022-TALENTS-01 financed by European Research Executive Agency (REA), IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program), and the European Lighthouse of AI for Sustainability [grant number 101120237-HORIZON-CL4-2022-HUMAN-02].


## Competing interests

The authors declare no competing interests.

## Data and materials availability

All data are available in the manuscript or the supplementary materials.

# References


1. S. Athey, "The Impact of Machine Learning on Economics" in *The Economics of Artificial Intelligence: An Agenda*, National Bureau of Economic Research conference report., A. Agrawal, J. Gans, A. Goldfarb, National Bureau of Economic Research, Eds. (The University of Chicago Press, 2019), pp. 507–547.
2. S. Athey, G. W. Imbens, Machine Learning Methods That Economists Should Know About. *Annu. Rev. Econ.* **11**, 685–725 (2019).
3. N. Jean, *et al.*, Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
4. D. Ahn, *et al.*, A human-machine collaborative approach measures economic development using satellite imagery. *Nat. Commun.* **14**, 6811 (2023).
5. G. Chi, H. Fang, S. Chatterjee, J. E. Blumenstock, Microestimates of wealth for all low- and middle-income countries. *Proc. Natl. Acad. Sci.* **119**, e2113658119 (2022).
6. J. V. Henderson, A. Storeygard, D. N. Weil, Measuring Economic Growth from Outer Space. *Am. Econ. Rev.* **102**, 994–1028 (2012).
7. C. Robinson, F. Hohman, B. Dilkina, A Deep Learning Approach for Population Estimation from Satellite Imagery in *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, (ACM, 2017), pp. 47–54.
8. F. R. Stevens, A. E. Gaughan, C. Linard, A. J. Tatem, Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLOS ONE* **10**, e0107042 (2015).
9. M. C. Hansen, *et al.*, High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* **342**, 850–853 (2013).
10. E. Rolf, *et al.*, A generalizable and accessible approach to machine learning with global satellite imagery. *Nat. Commun.* **12**, 4392 (2021).
11. M. Burke, A. Driscoll, D. B. Lobell, S. Ermon, Using satellite imagery to understand and promote sustainable development. *Science* **371**, eabe8628 (2021).
12. L. Yu, *et al.*, Meta-discoveries from a synthesis of satellite-based land-cover mapping research. *Int. J. Remote Sens.* **35**, 4573–4588 (2014).
13. C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann, The product space conditions the development of nations. *Science* **317**, 482–487 (2007).
14. F. L. Pinheiro, D. Hartmann, R. Boschma, C. A. Hidalgo, The time and frequency of unrelated diversification. *Res. Policy* 104323 (2021). https://doi.org/10.1016/j.respol.2021.104323.
15. S. Poncet, F. S. de Waldemar, Product Relatedness and Firm Exports in China. *World Bank Econ. Rev.* **29**, 579–605 (2015).
16. C. A. Hidalgo, Economic complexity theory and applications. *Nat. Rev. Phys.* **3**, 92–113 (2021).
17. J. L. van Zanden, B. van Leeuwen, Persistent but not consistent: The growth of national income in Holland 1347–1807. *Explor. Econ. Hist.* **49**, 119–130 (2012).
18. S. N. Broadberry, B. M. S. Campbell, A. Klein, M. Overton, B. van Leeuwen, *British economic growth, 1270-1870* (Cambridge University Press, 2015).
19. N. Palma, J. Reis, From Convergence to Divergence: Portuguese Economic Growth, 1527–1850. *J. Econ. Hist.* **79**, 477–506 (2019).
20. P. Malanima, The long decline of a leading economy: GDP in central and northern Italy, 1300-1913. *Eur. Rev. Econ. Hist.* **15**, 169–219 (2011).

21. U. Pfister, Economic Growth in Germany, 1500–1850. *J. Econ. Hist.* **82**, 1071–1107 (2022).
22. M. Malinowski, J. L. van Zanden, Income and its distribution in preindustrial Poland. *Cliometrica* **11**, 375–404 (2017).
23. C. Álvarez-Nogal, L. P. De La Escosura, The rise and fall of Spain (1270-1850). *Econ. Hist. Rev.* **66**, 1–37 (2013).
24. O. Krantz, Swedish GDP 1300-1560 : A Tentative Estimate. *Lund Pap. Econ. Hist. Gen. Issues* **152** (2017).
25. L. Ridolfi, Six Centuries of Real Wages in France from Louis IX to Napoleon III: 1250–1860. *J. Econ. Hist.* **79**, 589–627 (2019).
26. L. Schön, O. Krantz, New Swedish Historical National Accounts since the 16th Century in Constant and Current Prices. *Lund Pap. Econ. Hist. Gen. Issues* **140** (2015).
27. J. Bolt, J. L. van Zanden, Maddison style estimates of the evolution of the world economy. A new 2020 update. *Maddison-Proj. Work. Pap.* **WP-15** (2020).
28. J. Bolt, J. L. van Zanden, The Maddison Project: collaborative research on historical national accounts: The Maddison Project. *Econ. Hist. Rev.* **67**, 627–651 (2014).
29. C. Alvarez-Nogal, L. P. De La Escosura, The decline of Spain (1500-1850): conjectural estimates. *Eur. Rev. Econ. Hist.* **11**, 319–366 (2007).
30. K. Enflo, A. Missiaia, Regional GDP estimates for Sweden, 1571–1850. *Hist. Methods J. Quant. Interdiscip. Hist.* **51**, 115–137 (2018).
31. M. Laouenan, *et al.*, A cross-verified database of notable people, 3500BC-2018AD. *Sci. Data* **9**, 290 (2022).
32. A. Z. Yu, S. Ronen, K. Hu, T. Lu, C. A. Hidalgo, Pantheon 1.0, a manually verified dataset of globally famous biographies. *Sci. Data* **3**, 150075 (2016).
33. J. Mokyr, "Long-Term Economic Growth and the History of Technology" in *Handbook of Economic Growth*, (Elsevier, 2005), pp. 1113–1180.
34. U. Akcigit, J. Grigsby, T. Nicholas, The rise of Amercian ingenuity: Innovation and inventors of the Golden Age. *NBER Work. Pap.* **23047** (2017).
35. C. MacLaurin, P. Murdoch, *An account of Sir Isaac Newton's philosophical discoveries: in four book*, 2nd Ed. (Printed for A. Millar, 1750).
36. M. Schich, *et al.*, A network framework of cultural history. *Science* **345**, 558–562 (2014).
37. J. Mokyr, Mobility, Creativity, and Technological Development: David Hume, Immanuel Kant and the Economic Development of Europe in *Session on "Creativity and the Economy", German Association of Philosophy*, (2005).
38. J. O'Hagan, K. J. Borowiecki, Birth Location, Migration, and Clustering of Important Composers: Historical Patterns. *Hist. Methods J. Quant. Interdiscip. Hist.* **43**, 81–90 (2010).
39. P. Koch, V. Stojkoski, C. A. Hidalgo, The Role of Immigrants, Emigrants, and Locals in the Historical Formation of European Knowledge Agglomerations. *Reg. Stud.* (2023). https://doi.org/10.1080/00343404.2023.2275571.
40. C. M. Cipolla, The Diffusion of Innovations in Early Modern Europe. *Comp. Stud. Soc. Hist.* **14**, 46–52 (1972).
41. E. Miguelez, A. Morrison, Migrant Inventors as Agents of Technological Change. *J. Technol. Transf.* (2022). https://doi.org/10.1007/s10961-022-09927-z.
42. D. Bahar, P. Choudhury, H. Rapoport, Migrant inventors and the technological advantage of nations. *Res. Policy* **49**, 103947 (2020).

43. S. Bernstein, R. Diamond, A. Jiranaphawiboon, T. McQuade, B. Pousada, The Contribution of High-Skilled Immigrants to Innovation in the United States. *NBER Work. Pap. Ser.* w30797 (2022). https://doi.org/10.3386/w30797.
44. E. Hornung, Immigration and the Diffusion of Technology: The Huguenot Diaspora in Prussia. *Am. Econ. Rev.* **104**, 84–122 (2014).
45. I. Ganguli, Immigration and Ideas: What Did Russian Scientists "Bring" to the United States? *J. Labor Econ.* **33**, S257–S288 (2015).
46. D. Diodato, A. Morrison, S. Petralia, Migration and invention in the Age of Mass Migration. *J. Econ. Geogr.* **22**, 477–498 (2022).
47. K. J. Borowiecki, K. Graddy, Immigrant artists: Enrichment or displacement? *J. Econ. Behav. Organ.* **191**, 785–797 (2021).
48. K. J. Borowiecki, Are composers different? Historical evidence on conflict-induced migration (1816-1997). *Eur. Rev. Econ. Hist.* **16**, 270–291 (2012).
49. D. Acemoglu, S. Johnson, J. Robinson, The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth. *Am. Econ. Rev.* **95**, 546–579 (2005).
50. A. M. de Pleijt, J. L. van Zanden, Accounting for the "Little Divergence": What drove economic growth in pre-industrial Europe, 1300–1800? *Eur. Rev. Econ. Hist.* **20**, 387–409 (2016).
51. A. Henriques, N. Palma, Comparative European Institutions and the Little Divergence, 1385–1800. *J. Econ. Growth* (2022). https://doi.org/10.1007/s10887-022-09213-5.
52. M. Fochesato, Origins of Europe's north-south divide: Population changes, real wages and the 'little divergence' in early modern Europe. *Explor. Econ. Hist.* **70**, 91–131 (2018).
53. R. C. Allen, The Great Divergence in European Wages and Prices from the Middle Ages to the First World War. *Explor. Econ. Hist.* **38**, 411–447 (2001).
54. E. Buringh, The Population of European Cities from 700 to 2000: Social and Economic History. *Res. Data J. Humanit. Soc. Sci.* **6**, 1–18 (2021).
55. J. Baten, M. Blum, Why are you tall while others are short? Agricultural production and other proximate determinants of global heights. *Eur. Rev. Econ. Hist.* **18**, 144–165 (2014).
56. A. Rijpma, "A composite view of well-being since 1820" in *How Was Life?*, J. L. Van Zanden, J. Baten, M. Mira d'Ercole, A. Rijpma, M. P. Timmer, Eds. (OECD, 2014), pp. 249–269.
57. E. Buringh, B. M. S. Campbell, A. Rijpma, J. L. Van Zanden, Church building and the economy during Europe's 'Age of the Cathedrals', 700–1500 CE. *Explor. Econ. Hist.* **76**, 101316 (2020).
58. K. Enflo, M. Henning, L. Schön, "Swedish Regional GDP 1855–2000: Estimations and General Trends in the Swedish Regional System" in *Research in Economic History*, (Emerald Group Publishing, 2014), pp. 47–89.
59. N. Delefortrie, J. Morice, Les revenus départementaux en 1864 et en 1954. *Population* **15**, 721 (1960).
60. M. P. Squicciarini, N. Voigtländer, Human Capital and Industrialization: Evidence from the Age of Enlightenment *. *Q. J. Econ.* **130**, 1825–1883 (2015).
61. F. Geary, T. Stark, Regional GDP in the UK, 1861-1911: new estimates: Regional GDP. *Econ. Hist. Rev.* **68**, 123–144 (2015).
62. Office for National Statistics, Historical Regional GDP 1968 to 1970 and 1971 to 1996. Office for National Statistics. Deposited 2016.
63. E. Felice, The roots of a dual equilibrium: GDP, productivity, and structural change in the Italian regions in the long run (1871–2011). *Eur. Rev. Econ. Hist.* (2018). https://doi.org/10.1093/ereh/hey018.

64. M. Badia-Miró, J. Guilera, P. Lains, Reconstruction of the Regional GDP of Portugal, 1890 - 1980. *UB Econ. - Work. Pap.* **12/280** (2012).
65. E. Buyst, Reversal of Fortune in a Small, Open Economy: Regional GDP in Belgium, 1896-2000. *SSRN Electron. J.* (2009). https://doi.org/10.2139/ssrn.1586762.
66. Eurostat, Gross domestic product (GDP) at current market prices by NUTS 2 regions. Eurostat. Deposited 2023.
67. Office for National Statistics, Regional gross domestic product: all ITL regions. Office for National Statistics. Deposited 2022.
68. Bureau of Economic Analysis, Gross Domestic Product by Metropolitan Area. Bureau of Economic Analysis. Deposited 2018.
69. Statistics Canada, Metropolitan gross domestic product. Statistics Canada. Deposited 2014.
70. State Statistics Services Ukraine, Валовий регіональний продукт. Deposited 2013.
71. Belstat, Gross regional product at current prices. Deposited 2023.
72. Rosstat, Gross Regional Product at current basic prices per capita (1998-2019). Deposited 2020.
73. L. Prados De La Escosura, International Comparisons of Real Product, 1820–1990: An Alternative Data Set. *Explor. Econ. Hist.* **37**, 1–41 (2000).
74. M. J. Klasing, P. Milionis, Quantifying the evolution of world trade, 1870–1949. *J. Int. Econ.* **92**, 185–197 (2014).
75. K. Pomeranz, *The great divergence: China, Europe, and the making of the modern world economy*, First Princeton Classics Paperback Edition (Princeton University Press, 2021).
76. R. Inklaar, H. de Jong, J. Bolt, J. L. Van Zanden, Rebasing "Maddison": new income comparisons and the shape of long-run economic development. *Gron. Growth Dev. Cent. Res. Memo.* **174** (2018).
77. C. Hube, Bias in Wikipedia in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, (ACM Press, 2017), pp. 717–721.
78. M. Dittus, M. Graham, Mapping Wikipedia's Geolinguistic Contours. *Digit. Cult. Soc.* **5**, 147–164 (2019).
79. U. Pfeil, P. Zaphiris, C. S. Ang, Cultural Differences in Collaborative Authoring of Wikipedia. *J. Comput.-Mediat. Commun.* **12**, 88–113 (2006).
80. E. S. Callahan, S. C. Herring, Cultural bias in Wikipedia content on famous persons. *J. Am. Soc. Inf. Sci. Technol.* **62**, 1899–1915 (2011).
81. A. Halavais, D. Lackaff, An Analysis of Topical Coverage of Wikipedia. *J. Comput.-Mediat. Commun.* **13**, 429–440 (2008).
82. T. Chesney, An empirical examination of Wikipedia's credibility. *First Monday* **11** (2006).
83. J. Giles, Internet encyclopaedias go head to head. *Nature* **438**, 900–901 (2005).
84. M. Serafinelli, G. Tabellini, Creativity over time and space: A historical analysis of European cities. *J. Econ. Growth* **27**, 1–43 (2022).
85. D. De La Croix, O. Licandro, The longevity of famous people from Hammurabi to Einstein. *J. Econ. Growth* **20**, 263–303 (2015).
86. C. A. Hidalgo, R. Hausmann, The building blocks of economic complexity. *PNAS* **106**, 10570–10575 (2009).
87. P.-A. Balland, *et al.*, The new paradigm of economic complexity. *Res. Policy* **51**, 104450 (2022).
88. V. Stojkoski, Z. Utkovski, L. Kocarev, The Impact of Services on Economic Complexity: Service Sophistication as Route for Economic Growth. *PLOS ONE* **11**, e0161633 (2016).

89. P. Koch, Economic complexity and growth: Can value-added exports better explain the link? *Econ. Lett.* **198**, 109682 (2021).
90. G. Domini, Patterns of specialization and economic complexity through the lens of universal exhibitions, 1855-1900. *Explor. Econ. Hist.* 101421 (2022). https://doi.org/10.1016/j.eeh.2021.101421.
91. R. Hausmann, *et al.*, *The Atlas of economic complexity: Mapping paths to prosperity* (Center for International Development, Harvard University and Harvard Kennedy School and Macro Connections, MIT and Massachusetts Institute of Technology, 2011).
92. V. Stojkoski, P. Koch, C. A. Hidalgo, Multidimensional economic complexity and inclusive green growth. *Commun. Earth Environ.* **4**, 130 (2023).
93. I. M. Weber, G. Semieniuk, T. Westland, J. Liang, What You Exported Matters: Persistence in Productive Capabilities across Two Eras of Globalization. *UMass Amherst Econ. Dep. Work. Pap. Ser.* **2021–02** (2021).
94. D. Hartmann, M. R. Guevara, C. Jara-Figueroa, M. Aristarán, C. A. Hidalgo, Linking Economic Complexity, Institutions, and Income Inequality. *World Dev.* **93**, 75–93 (2017).
95. A. Lapatinas, The effect of the Internet on economic sophistication: An empirical analysis. *Econ. Lett.* **174**, 35–38 (2019).
96. J. P. Romero, C. Gramkow, Economic complexity and greenhouse gas emissions. *World Dev.* **139**, 105317 (2021).
97. H. Zou, T. Hastie, Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
98. P. Bairoch, J. Batou, P. Chèvre, *La population des villes européennes de 800 à 1850* (Librairie Droz, 1988).
99. J. E. Dittmar, Information Technology and Economic Change: The Impact of The Printing Press. *Q. J. Econ.* **126**, 1133–1172 (2011).
100. J. B. De Long, A. Shleifer, Princes and Merchants: European City Growth before the Industrial Revolution. *J. Law Econ.* **36**, 671–702 (1993).
101. J. Mokyr, C. Ó Gráda, Height and Health in the United Kingdom 1815–1860: Evidence from the East India Company Army. *Explor. Econ. Hist.* **33**, 141–168 (1996).
102. H. J. Brinkman, J. W. Drukker, B. Slot, Height and income: A new method for the estimation of historical national income series. *Explor. Econ. Hist.* **25**, 227–264 (1988).
103. R. H. Steckel, Height and Per Capita Income. *Hist. Methods J. Quant. Interdiscip. Hist.* **16**, 1–7 (1983).
104. J. C. Smith, *Nuremberg, a Renaissance city, 1500-1618*, 1st ed (Published for the Archer M. Huntington Art Gallery, the University of Texas at Austin [by the] University of Texas Press, 1983).
105. J. C. Smith, Netherlandish Artists and Art in Renaissance Nuremberg. *Simiolus Neth. Q. Hist. Art* **20**, 153 (1990).
106. S. O. Becker, L. Woessmann, Was Weber Wrong? A Human Capital Theory of Protestant Economic History. *Q. J. Econ.* **124**, 531–596 (2009).
107. S. O. Becker, S. Pfaff, J. Rubin, Causes and consequences of the Protestant Reformation. *Explor. Econ. Hist.* **62**, 1–25 (2016).
108. D. Cantoni, The Economic Effects of the Protestant Reformation: Testing the Weber Hypothesis in the German Lands. *J. Eur. Econ. Assoc.* **13**, 561–598 (2015).
109. R. C. Nash, Urbanization in the Colonial South: Charleston, South Carolina, as a Case Study. *J. Urban Hist.* **19**, 3–29 (1992).
110. A. L. Slap, F. Towers, D. R. Goldfield, Eds., *Confederate cities: the urban South during the Civil War era* (The University of Chicago Press, 2015).

111. A. S. Pereira, The Opportunity of a Disaster: The Economic Impact of the 1755 Lisbon Earthquake. *J. Econ. Hist.* **69**, 466 (2009).
112. L. Shapley, “A value for n-person games” in *Contributions to the Theory of Games (AM-28). Volume 2*, (Princeton University Press, 1953), pp. 307–317.
113. S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **30** (2017).
114. B. Rozemberczki, *et al.*, The Shapley Value in Machine Learning. (2022). https://doi.org/10.48550/ARXIV.2202.05594.
115. G. H. Golub, C. Reinsch, “Singular Value Decomposition and Least Squares Solutions” in *Handbook for Automatic Computation*, F. L. Bauer, *et al.*, Eds. (Springer Berlin Heidelberg, 1971), pp. 134–151.

# DISCUSSION

Machine learning methods became a crucial methodological resource in economic over the past decade (1–3). As this thesis shows, machine learning is also transforming research in economic history and helps us better understand certain aspects of the past.

The first chapter showed that machine learning applications in economic history broadly fall into three clusters. First, text recognition and natural language processing are applied to digitize archival resources and make them available for empirical research. This ranges from patents (4–6) to occupation classifications (7) and census data (8–10). Second, unsupervised machine learning models are used to derive new variables that help us understand historical developments. This involves text-based methods such as structural topic modelling (11, 12), but also dimensionality reduction techniques such as economic complexity (13, 14) and recommender systems such as relatedness measures (15). Third, recent efforts generate new data building upon supervised machine learning models (16–19).

The second and third chapter of this thesis contribute to the latter two clusters.

In the second chapter, we used unsupervised machine learning models—that is, measures of relatedness (20, 21)—to explore the role of immigrants, emigrants, and locals in shaping the specializations of European agglomerations over the past 1,000 years. We used biographic data on more than 22,000 famous individuals—sculptors, composers, politicians, chemists, etc.—living in Europe between the years 1000 and 2000 to explore how the knowledge of migrants and locals explains the probability that a region enters or exits an activity. Our findings showed that migrants play a crucial role in the historical geography of knowledge. Specifically, we found that the probability that a European region enters a new activity grows with the presence of immigrants with knowledge on that activity and related activities. Put differently, the probability that a region begets famous mathematicians grows with an excess immigration of mathematicians and with immigrants from related fields, such as physics or chemistry. These findings advance our understanding of the evolution of European agglomerations over the past millennium and of the role of migrants and locals therein.

In the third chapter, we explored whether data on the biographies of historical figures can be used to create models of historical GDP per capita levels for countries and regions in Europe and North America for the past 700 years and estimate their historical GDP per capita. Leveraging information on more than 563k historical figures recorded across multiple languages in Wikipedia (22, 23) we were able to construct a supervised machine learning model that makes relatively accurate predictions. In an out-of-sample test, this model predicts the GDP per capita of European and North American countries and regions with an $R^2$=90.1% and a

mean absolute error of 22.6% of the GDP per capita observed during that time period. This data does not only allow for investigating 700 years of cross-country differences in economic development, but also for comparing the development of different regions in Europe (Milan, Montpellier, Paris, London, etc.) with metropolitan areas in North America (New York, Boston, Toronto etc.).

Together, these chapters show that machine learning methods can help massively in making historical data available for quantitative analyses and even augmenting its availability beyond existing resources. Also, machine learning enables us to explore historical research questions that were relatively unexplored, such as the role of migrants and locals in shaping the cultural specializations of European cities.

Still, I believe we are just at the beginning of seeing machine learning methods impact research in economic history. There are several promising avenues for future research that reduce limitations of current contributions and open completely new research directions.

First, machine learning methods will keep improving. When matching observations across datasets, for instance, machine learning methods still produce a substantial amount of false positives, which can bias empirical results building upon it (24). More generally, I believe that machine learning methods will become more readily available to researchers, and more reliable.

Second, the frontier in ML and artificial intelligence methods will impact research in economics and economic history. Consider Large Language Models that took the world by storm after 2022. These models enable a completely different and more efficient approach to handling large amounts of unstructured text data.

As this thesis showed, a specific text data source can be highly valuable in economic history research: biographies of famous individuals. One limitation of current studies, including the second and third chapter of this thesis, is the lack of granular information on migration patterns of famous individuals. Up to now, places of birth and death are typically used as a proxy for migration (22, 25, 26). While this is a solid proxy (as the supplementary materials to Chapter 2 showed), famous individuals have been remarkably mobile (27). Einstein was born in Ulm in Germany and died in Princeton but lived in several cities in the German-speaking world in the meantime. All this information is available in encyclopedias as unstructured text data. Recent advances such as LLMs can help extract this information in a structured manner. Having more detailed data on where famous individuals lived and when could provide a better analytical basis to explore the evolution of agglomerations (Chapter 2) and augment the availability of historical GDP per capita data (Chapter 3).

Lastly, I believe that the use of machine learning models to generate new data or improve existing data is the most promising future avenue of research. Recovering the location of lost cities (17) or augmenting the availability of historical GDP per capita estimates (Chapter 3) are important milestones in better understanding the past. The impact of these approaches, however, crucially depends on the quantity and quality of the available data. With better methods for pre-processing and the adoption of new technologies such as Large Language Models, the quantity and quality of input data will increase substantially.

# References


1. S. Athey, "The Impact of Machine Learning on Economics" in *The Economics of Artificial Intelligence: An Agenda*, National Bureau of Economic Research conference report., A. Agrawal, J. Gans, A. Goldfarb, National Bureau of Economic Research, Eds. (The University of Chicago Press, 2019), pp. 507–547.
2. S. Athey, M. Bayati, N. Doudchenko, G. Imbens, K. Khosravi, Matrix Completion Methods for Causal Panel Data Models. *J. Am. Stat. Assoc.* **116**, 1716–1730 (2021).
3. S. Athey, G. W. Imbens, Machine Learning Methods That Economists Should Know About. *Annu. Rev. Econ.* **11**, 685–725 (2019).
4. D. Diodato, A. Morrison, S. Petralia, Migration and invention in the Age of Mass Migration. *J. Econ. Geogr.* **22**, 477–498 (2022).
5. S. Petralia, P.-A. Balland, D. L. Rigby, Unveiling the geography of historical patents in the United States from 1836 to 1975. *Sci. Data* **3**, 160074 (2016).
6. F. Van Der Wouden, A history of collaboration in US invention: changing patterns of co-invention, complexity and geography. *Ind. Corp. Change* **29**, 599–619 (2020).
7. L. Nedelkoska, *et al.*, Eight Decades of Changes in Occupational Tasks, Computerization and the Gender Pay Gap. [Preprint] (2021). Available at: https://www.ifo.de/sites/default/files/events/2022/pillars22-Nedelkoska.pdf.
8. R. Abramitzky, L. Boustan, K. Eriksson, J. Feigenbaum, S. Pérez, Automated Linking of Historical Data. *J. Econ. Lit.* **59**, 865–918 (2021).
9. R. Hornbeck, S. H.-M. Hsu, A. Humlum, M. Rotemberg, Technological Stickiness: Switching and Entry in the Long Transition from Water to Steam Power. [Preprint] (2023). Available at: https://www.stern.nyu.edu/sites/default/files/2023-10/mrotembreg_nyu_macro_steam.pdf.
10. J. Price, K. Buckles, J. Van Leeuwen, I. Riley, Combining family history and machine learning to link historical records: The Census Tree data set. *Explor. Econ. Hist.* **80**, 101391 (2021).
11. P. Grajzl, P. Murrell, A machine-learning history of English caselaw and legal ideas prior to the Industrial Revolution I: generating and interpreting the estimates. *J. Institutional Econ.* **17**, 1–19 (2021).
12. P. Grajzl, P. Murrell, A machine-learning history of English caselaw and legal ideas prior to the Industrial Revolution II: applications. *J. Institutional Econ.* **17**, 201–216 (2021).

13. G. Domini, Patterns of specialization and economic complexity through the lens of universal exhibitions, 1855-1900. *Explor. Econ. Hist.* 101421 (2022). https://doi.org/10.1016/j.eeh.2021.101421.
14. I. M. Weber, G. Semieniuk, T. Westland, J. Liang, What You Exported Matters: Persistence in Productive Capabilities across Two Eras of Globalization. *UMass Amherst Econ. Dep. Work. Pap. Ser.* **2021–02** (2021).
15. P. Koch, V. Stojkoski, C. A. Hidalgo, The Role of Immigrants, Emigrants, and Locals in the Historical Formation of European Knowledge Agglomerations. *Reg. Stud.* 1–15 (2023). https://doi.org/10.1080/00343404.2023.2275571.
16. P. Koch, V. Stojkoski, C. A. Hidalgo, Augmenting the availability of historical GDP per capita estimates through machine learning. [Preprint] (2024). Available at: https://static1.squarespace.com/static/646fc00fbb5b0c3b5be4f496/t/6684044a1fcabb0c6d1e9889/1719927888376/manuscript_historicalGDPpc_20240702.pdf.
17. G. Barjamovic, T. Chaney, K. Coşar, A. Hortaçsu, Trade, Merchants, and the Lost Cities of the Bronze Age. *Q. J. Econ.* **134**, 1455–1503 (2019).
18. E. Chaney, Modern Library Holdings and Historic City Growth. [Preprint] (2022). Available at: https://www.tse-fr.eu/sites/default/files/TSE/documents/sem2022/bid/chaney.pdf.
19. M. Saavedra, T. Twinam, A machine learning approach to improving occupational income scores. *Explor. Econ. Hist.* **75**, 101304 (2020).
20. C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann, The product space conditions the development of nations. *Science* **317**, 482–487 (2007).
21. C. A. Hidalgo, *et al.*, "The Principle of Relatedness" in *Unifying Themes in Complex Systems IX*, Springer Proceedings in Complexity., A. J. Morales, C. Gershenson, D. Braha, A. A. Minai, Y. Bar-Yam, Eds. (Springer International Publishing, 2018), pp. 451–457.
22. M. Laouenan, *et al.*, A cross-verified database of notable people, 3500BC-2018AD. *Sci. Data* **9**, 290 (2022).
23. A. Z. Yu, S. Ronen, K. Hu, T. Lu, C. A. Hidalgo, Pantheon 1.0, a manually verified dataset of globally famous biographies. *Sci. Data* **3**, 150075 (2016).
24. M. Bailey, C. Cole, M. Henderson, C. Massey, "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data" (National Bureau of Economic Research, 2019).
25. M. Schich, *et al.*, A network framework of cultural history. *Science* **345**, 558–562 (2014).
26. M. Serafinelli, G. Tabellini, Creativity over time and space: A historical analysis of European cities. *J. Econ. Growth* **27**, 1–43 (2022).
27. J. Mokyr, Mobility, Creativity, and Technological Development: David Hume, Immanuel Kant and the Economic Development of Europe in *Session on "Creativity and the Economy", German Association of Philosophy*, (2005).

# SUPPLEMENTARY MATERIALS
# FOR CHAPTER 2

# Content

# 1. Data

## 1.1. Pantheon

The main data source for our analysis is the 2020 version of the Pantheon dataset (Yu et al., 2016), which is publicly available at pantheon.world. It contains information on more than 88,000 famous individuals with more than 15 language editions in Wikipedia worldwide. We restrict our sample to the years 1000 to 2000, since the number of observations in the overall dataset increases after the year 1000 and becomes less volatile (see Figure S3). Also, as described in the main manuscript, we focus on continental Europe and, thus, only include individuals who are born or have died in Europe. The reasoning behind this restriction is the need to have an as comprehensive picture of the structure of famous individuals in a region as possible. Due to an arguable Western bias in Wikipedia and the selected time horizon, we restrict our sample to Europe. Overall, this reduces our sample to 22,847 individuals.

We follow the occupation taxonomy by Yu et al. (2016), which differentiates in total between 101 occupations of 27 categories and 8 broad categories. Table S2 describes the taxonomy and displays the number of famous individuals born or died in Europe between 1000 and 2000 with the respective occupation. Politicians (5,233), writers (2,817) and painters (1,126) are the most common occupations of famous figures in the past millennium.

Occupations are assigned to individuals based on the occupation that made them famous. For instance, Marie Curie is considered a physicist in our dataset, since she won the Nobel Prize in physics prior to her Nobel Prize in chemistry. Angela Merkel is considered a politician, despite her academic career in chemistry. A more detailed description of this approach is given in Yu et al. (2016).

A consistent occupation classification is a key element for our analysis, since we want to describe the geography of knowledge based on this classification. In fact, more comprehensive data sources for notable people would be available, such as Freebase from Google or a very recently published database (Laouenan et al., 2022), which contains information on 2.29 million notable individuals. Unfortunately, these data sources are not sufficiently consistent with respect to their occupation classification. For example, the database by Laouenan et al. (2022) distinguishes between almost 5,000 occupations, but these are not unique. Nonetheless, these data sources are very promising avenues for future research, potentially enabling the analysis of the historical geography of knowledge based on notable figures beyond Europe.

The number of observations in or dataset increases with time (Figure S3). While we have data on 284 individuals born in the 11$^{th}$ century, this number increases to 7,483 in the 20$^{th}$ century

(see Table S3). Due to this imbalance, we perform robustness checks for period subsamples in section 3.4.3.

To obtain a comprehensive picture of the knowledge of individuals living in a location given the unbalanced sample (Table S3), the period of observation in our study is centuries. Specifically, we assign individuals to centuries based on their year of birth. A famous individual is, for instance, assigned to the 17$^{th}$ century if he or she is born between 1600 and 1699. Splitting the sample into smaller time periods such as decades or half-centuries would prohibit us from estimating measures of specialization or relatedness in all periods due to small numbers of observation.

We do not observe the full migration trajectory of individuals, which is why we use places of birth and death as a proxy for migration (see section 1.3). Interestingly, migration is very common among notable individuals in the dataset. 75.1 percent of individuals in the dataset die in a different region they are born in. Also, migration among famous individuals has become more prevalent over time. While in the 11$^{th}$ century 31.3 percent of individuals died in a different region than they were born in, this is only the case for 20.8 percent in the 19$^{th}$ century (see Table S3 and Figure 1e in the main text).

Individuals are assigned to centuries based solely on their year of birth. That is, a famous person who is born in the 18$^{th}$ century in Brussels and has later died in Paris is considered a local in Brussels and an immigrant in Paris in the 18$^{th}$ century. Even if the individual dies in the 19$^{th}$ century, the person is assigned an immigrant in the 18$^{th}$ century. We choose this approach, since we do not have information on the time of migration. It may be that the considered individual moved to Paris in the later stages of his or her life, but it may also be the case that the migration took place as a child. We do not believe that this definition is problematic in our analysis, given the lag structure in our regression model and the length of the chosen period of observation, i.e. centuries.

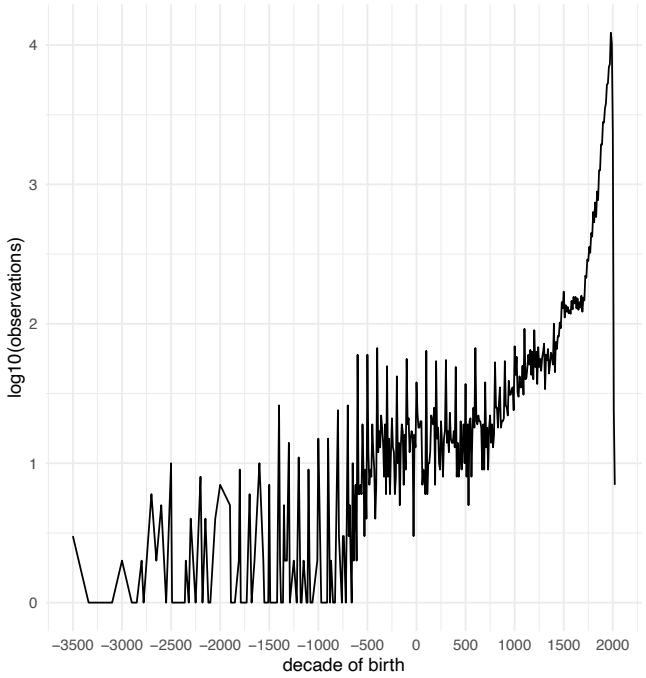



***Figure S3***. *Number of observations in the overall Pantheon dataset (Yu et al., 2016) by decade of birth*

***Table S2***. *Occupation taxonomy following Yu et al. (2016) and number of individuals born and/or deceased in Europe between the years 1000 and 2000*

| Broad Category | Category | Occupation | Obs. abs. | Obs. in % |
|---|---|---|---|---|
| Arts | Dance | DANCER | 46 | 0,2% |
| | Design | ARCHITECT | 300 | 1,3% |
| | | DESIGNER | 42 | 0,2% |
| | | COMIC ARTIST | 25 | 0,1% |
| | | FASHION DESIGNER | 14 | 0,1% |
| | Film and Theatre | ACTOR | 969 | 4,2% |
| | | FILM DIRECTOR | 421 | 1,8% |
| | Fine Arts | PAINTER | 1126 | 4,9% |
| | | SCULPTOR | 122 | 0,5% |
| | | PHOTOGRAPHER | 47 | 0,2% |
| | | ARTIST | 33 | 0,1% |
| | Music | COMPOSER | 889 | 3,9% |
| | | MUSICIAN | 390 | 1,7% |
| | | SINGER | 352 | 1,5% |
| | | CONDUCTOR | 64 | 0,3% |
| Business & Law | Business | BUSINESSPERSON | 158 | 0,7% |
| | | PRODUCER | 13 | 0,1% |
| | Law | LAWYER | 24 | 0,1% |
| Exploration | Explorers | EXPLORER | 295 | 1,3% |
| | | ASTRONAUT | 41 | 0,2% |
| Humanities | History | HISTORIAN | 184 | 0,8% |
| | Language | WRITER | 2817 | 12,4% |
| | | LINGUIST | 117 | 0,5% |
| | | JOURNALIST | 48 | 0,2% |
| | Philosophy | PHILOSOPHER | 563 | 2,5% |
| Institutions | Government | POLITICIAN | 5233 | 22,9% |
| | | NOBLEMAN | 551 | 2,4% |
| | | DIPLOMAT | 24 | 0,1% |
| | | PUBLIC WORKER | 8 | 0,0% |
| | Military | MILITARY PERSONNEL | 917 | 4,0% |
| | | PILOT | 33 | 0,1% |
| | Religion | RELIGIOUS FIGURE | 811 | 3,6% |
| Public Figure | Activism | SOCIAL ACTIVIST | 197 | 0,9% |
| | Media Personality | CELEBRITY | 31 | 0,1% |
| | | PRESENTER | 8 | 0,0% |
| | | MODEL | 8 | 0,0% |
| | Outlaws | EXTREMIST | 69 | 0,3% |
| | | OCCULTIST | 31 | 0,1% |
| | | PIRATE | 20 | 0,1% |
| | | MAFIOSO | 12 | 0,1% |

| Broad Category | Category | Occupation | Obs. abs. | Obs. in % |
|---|---|---|---|---|
| Science & Technology | Computer Science | COMPUTER SCIENTIST | 21 | 0,1% |
| | Engineering | ENGINEER | 208 | 0,9% |
| | Invention | INVENTOR | 205 | 0,9% |
| | Math | MATHEMATICIAN | 549 | 2,4% |
| | | STATISTICIAN | 6 | 0,0% |
| | Medicine | PHYSICIAN | 331 | 1,5% |
| | Natural Sciences | BIOLOGIST | 600 | 2,6% |
| | | PHYSICIST | 418 | 1,8% |
| | | CHEMIST | 330 | 1,4% |
| | | ASTRONOMER | 285 | 1,2% |
| | | ARCHAEOLOGIST | 79 | 0,3% |
| | | GEOLOGIST | 46 | 0,2% |
| | Social Sciences | ECONOMIST | 128 | 0,6% |
| | | PSYCHOLOGIST | 98 | 0,4% |
| | | GEOGRAPHER | 44 | 0,2% |
| | | ANTHROPOLOGIST | 36 | 0,2% |
| | | SOCIOLOGIST | 26 | 0,1% |
| | | POLITICAL SCIENTIST | 10 | 0,0% |
| Sports | Individual Sports | ATHLETE | 389 | 1,7% |
| | | RACING DRIVER | 260 | 1,1% |
| | | CYCLIST | 136 | 0,6% |
| | | CHESS PLAYER | 123 | 0,5% |
| | | TENNIS PLAYER | 79 | 0,3% |
| | | SKATER | 47 | 0,2% |
| | | WRESTLER | 43 | 0,2% |
| | | FENCER | 42 | 0,2% |
| | | BOXER | 39 | 0,2% |
| | | GYMNAST | 37 | 0,2% |
| | | SKIER | 37 | 0,2% |
| | | SWIMMER | 25 | 0,1% |
| | | MOUNTAINEER | 23 | 0,1% |
| | | TABLE TENNIS PLAYER | 6 | 0,0% |
| | Team Sports | SOCCER PLAYER | 963 | 4,2% |
| | | COACH | 32 | 0,1% |
| | | HOCKEY PLAYER | 22 | 0,1% |
| | | BASKETBALL PLAYER | 18 | 0,1% |
| | | HANDBALL PLAYER | 13 | 0,1% |

***Table S3**. Number of famous individuals and share of locals by century*

| Period | No. of observations | Share of individuals that died in the same region they were born in |
|---|---|---|
| 11th century | 284 | 31.3% |
| 12th century | 339 | 27.7% |
| 13th century | 377 | 32.1% |
| 14th century | 422 | 34.6% |
| 15th century | 784 | 31.6% |
| 16th century | 1,109 | 29.9% |
| 17th century | 1,231 | 34.6% |
| 18th century | 2,516 | 27.9% |
| 19th century | 8,301 | 20.8% |
| 20th century | 7,483 | 24.0% |

## 1.2. Administrative regions

For aggregation purposes, we assign individuals to regions based on their geocoded places of birth and death. We use NUTS-2 regions for countries in the European Union and the European Free Trade Association. For other countries in Europe, we use administrative regions of comparable size. Specifically, these are oblasts for Russia, Ukraine and Belarus, federal entities for Bosnia and Herzegovina, and the whole country for Kosovo and Moldova (see Figure S4). Shape files are publicly available online for NUTS regions (see e.g. ec.europa.eu) as well as for administrative regions of other countries (see e.g. gadm.org).

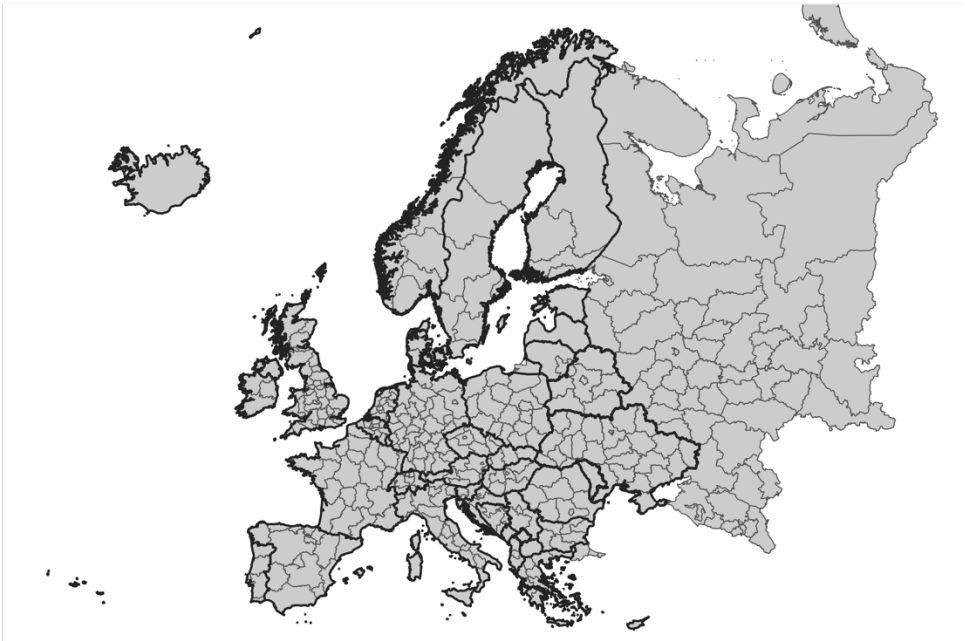


***Figure S4**. Administrative regions applied in the analysis. Bold lines mark country borders.*

## 1.3. Using places of birth and death as proxy for migration

We use places of birth and death as a proxy for migration, following the literature using similar data to describe migration movements (Laouenan et al., 2022; Schich et al., 2014; Serafinelli & Tabellini, 2022). Figures 1c and 1d in the main manuscript show the network of migration based on this proxy.

We check whether this proxy is, in fact, meaningful by randomly drawing ~200 individuals from the dataset, paying attention to representativeness across periods. We read the Wikipedia article for each famous individual to determine whether a relation to the place of death exists, which would qualify as migration. We differentiate between (1) having any relation to the place of death (i.e. living there for a considerable amount of time, having noteworthy social connections with multiple visits there, or, in case of politicians and noblemen, reigning over the region) and (2) having a major relation to the place of death. The latter is the case if the place was one of the individual's main places of living, if the famous individual taught at a university there etc.

We find that in 181 out of 202 cases ($\hat{p} = 0.896$, 95% CI: [0.854, 0.938]), the famous individual had a relation to his or her place of death. Hence, only in 10% of observations the place of death is arbitrary. Also, we find that in 151 out of 202 cases ($\hat{p} = 0.748$, 95% CI: [0.688, 0.807]), the famous individual had a major relation to his or her place of death. These results indicate that using place of birth and death as a proxy for migration is a valid approach. The sampled data is available upon request.

It is important to note that we do not claim that the place of death is the only and most relevant place of impact. Famous individuals, who seem to be highly mobile (see Table S3), are likely to stay at multiple cities during their lifetime. These stays, however, are not random. Famous individuals tend to spend time at places, where several individuals with the same specialization are already staying. This is a result of the negative binomial regression used to estimate the expected number of migrants and locals in Table S8.

Due to the observation that migration follows previous specialization patterns, we argue that if the estimates we find for the role of immigrants were affected by using place of birth and death as a proxy for migration and not observing the full migration trajectory, they would tend to be downward biased rather than upward biased.

## 1.4. Population

We augment our analysis with publicly available population data on more than 2,000 European cities going back to 700 AD (Bairoch et al., 1988; Buringh, 2021). We use the coordinates

provided in the dataset to assign cities to regions. This enables us to aggregate the population data by region.

# 2. Methods

## 2.1. Information entropy

Due to migration flows towards cities and the tendency of agglomeration, places of death are spatially more concentrated than places of birth. Figure S5 shows the number of births and deaths of famous individuals for the ten most populated regions in the 19th century. For example, 416 famous individuals were born in Paris in the 19th century, but 934 died there.

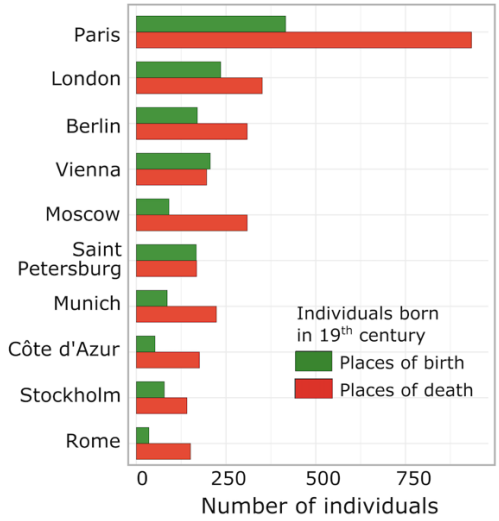



***Figure S5***. *Number of births and deaths of famous individuals born in the 19th century in the ten most common regions of death.*

We quantify the amount of spatial concentration by calculating the effective number of places of birth and places of death using information entropy.

Let $N_{i,t}$ denote the number of famous individuals in region $i$ and century $t$, then entropy $H$ is given by

$$H_t = -\sum_{i=1}^{I} \frac{N_{i,t}}{\sum_i N_{i,t}} log_2 \left( \frac{N_{i,t}}{\sum_i N_{i,t}} \right) \quad \text{(S1)}$$

Intuitively, $H_t$ is the average number of minimum yes/no questions one has to ask to guess a famous individual's region of birth or death.

The effective number of places is then given by $2^{H_t}$. This measure gives the number of places as if they were equally common. The higher the effective number of places, the lower the spatial concentration.

Figure 1f in the main manuscript plots the effective number of places of birth and places of death per century.

## 2.2. Adjacency matrices

To calculate the relatedness density, we transform our dataset into binary specialization matrices per century. We define that a location is specialized in an occupation if it exhibits a larger number of famous biographies in the respective occupation than expected. As described in the main manuscript, we employ two different approaches to the expected number of biographies. The first one is a naïve "bins and balls" model and identical to the Revealed Comparative Advantage or Location Quotient. The second approach consists of estimating the expected number of immigrants, emigrants and locals in a negative binomial regression model, taking local factors into account.

Then, we create specialization matrices based on immigrants (born somewhere else, but died here), emigration (born here, but died somewhere else) and locals (born here). We define the matrix $M_{ik}^{j}$ for $j = \{immi, emi, births\}$ as

$$M_{ik}^{j} = \begin{cases} 1 & if \ \frac{N_{ik,t}^{j}}{\widehat{N}_{ik,t}^{j}} \geq 1 \\ 0 & otherwise \end{cases} \quad \text{(S2)}$$

Prior to the calculation of the matrix $M_{ik}^{j}$ we remove regions and occupations with very few observations, since they can distort the specialization matrix. Specifically, we remove regions and occupations with not more than 5 famous individuals in a century, i.e. $\sum_k N_{ik}^{j} \leq 5$ and $\sum_i N_{ik}^{j} \leq 5$, respectively. For the 11th to 15th century, we employ a less restrictive cutoff, i.e. $\sum_k N_{ik}^{j} \leq 3$ and $\sum_i N_{ik}^{j} \leq 3$, due to fewer observations. Additionally, we remove individuals with the occupation "companion".

Sorting these specialization matrices by diversity and ubiquity reveals their nested structure (see Fig. 2a-c in the main manuscript).

## 2.3. The related knowledge of locals

We define locals as famous individuals who were born in a region, no matter if they died there or elsewhere. We use this definition because of the large share of migrants among famous individuals (see Table S3), which would reduce our number of observations drastically if we defined locals as individuals who were born and died in the same place. Here, we show that the relatedness density based on all famous individuals born is a valid proxy for the related

knowledge of individuals that have been born and died in the same region, after controlling for the related knowledge of emigrants.

To show this, we create a measure of relatedness for locals (born and died here) in analogy to the other relatedness measures based on the naïve model of the expected number:

$$M_{ik}^{locals} = \begin{cases} 1 & if \ \frac{N_{ik}^{locals} / \sum_k N_{ik}^{locals}}{\sum_i N_{ik}^{locals} / \sum_{i,k} N_{ik}^{locals}} \geq 1 \\ 0 & otherwise \end{cases}$$

$$\varphi_{kk\prime,t}^{locals} = \frac{\sum_i M_{ik,t}^{locals} M_{ik\prime,t}^{locals}}{max\left(\sum_i M_{ik,t}^{locals}, \sum_i M_{ik\prime,t}^{locals}\right)} \quad \text{(S3)}$$

$$\omega_{ik,t}^{locals} = \frac{\sum_{k'} M_{ik\prime,t}^{locals} \varphi_{kk\prime,t}^{locals}}{\sum_{k'} \varphi_{kk\prime,t}^{locals}}$$

Then, we estimate the following linear regression:

$$\omega_{ik,t}^{locals} = \alpha_1 \omega_{ik,t}^{births} + \alpha_2 \omega_{ik,t}^{emi} + \gamma_i + \delta_t + \varepsilon_{ik} \quad \text{(S4)}$$

Figure S6 shows the correlation between the fitted values based on the regression and the real values of $\omega_{ik,t}^{locals}$. The correlation between the real and fitted values is high ($R^2$=0.65), indicating that $\omega_{ik,t}^{births}$ controlling for $\omega_{ik,t}^{emi}$ is a valid proxy for $\omega_{ik,t}^{locals}$. This result is also robust for including all covariates of the logistic regression models described in section 3.1.

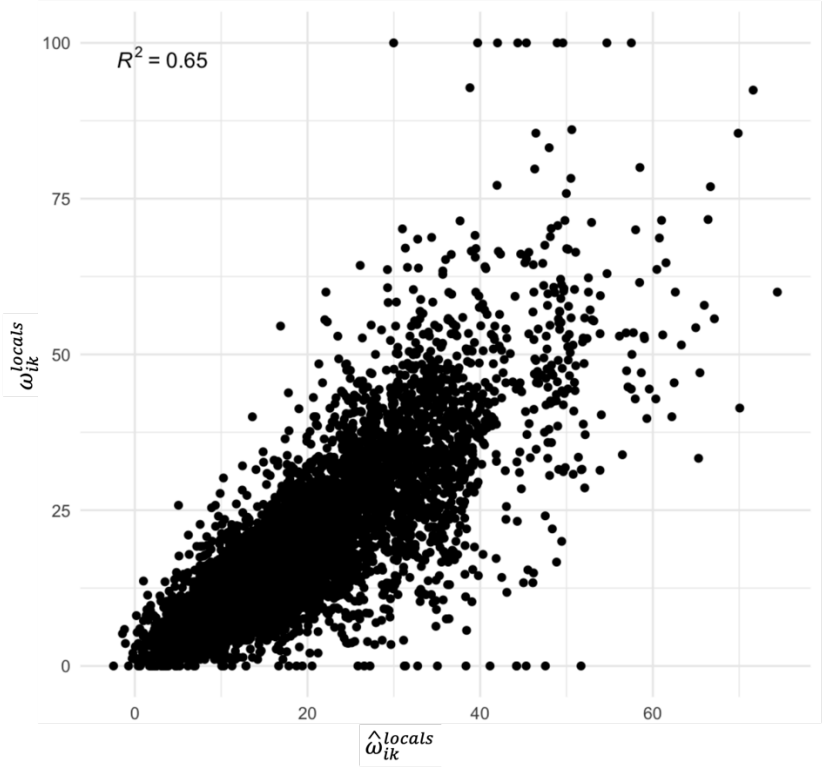



***Figure S6**. Correlation between fitted and real values for $\omega_{ik,t}^{locals}$*

## 2.4. Spatial lags

To control for other means of knowledge diffusion across space than migration, we create spatial lags. Specifically, we differentiate between the availability of knowledge in the same activity (do geographically proximate regions have a specialization in that activity?), and related activities (do geographically proximate regions have specializations in related activities?).

To do so, we measure the distance between all the administrative regions depicted in Figure S4 (technically, their centroids) and transform them into a proximity matrix $W_{i\prime i}$. Let $d_{i\prime i}$ denote the distance between regions $i'$ and $i$. Then,

$$W_{i\prime i} = \begin{cases} {}^{1}/_{d_{i\prime i}} & if\ i' \neq i \\ 0 & otherwise \end{cases} \quad \text{(S5)}$$

We then define the spatial lag with respect to specializations in the same activity as the region's average proximity to regions with a specialization in that activity:

$$\rho^{M}_{ik,t} = \frac{\sum_i W_{i'i} M^{births}_{ik,t}}{\sum_i W_{i'i}} \quad \text{(S6)}$$

Similarly, we define the spatial lag with respect to relatedness:

$$\rho^{\omega}_{ik,t} = \frac{\sum_i W_{i\prime i} \omega^{births}_{ik,t}}{\sum_i W_{i\prime i}} \quad \text{(S7)}$$

## 2.5. Elaboration on proximity measures

Measures of proximity capture the combined presence of multiple factors that may be contributing to the colocation of two activities. We create separate measures of proximity for immigrants, emigrants and locals (Eq. 8 in the main manuscript), since the factors driving the colocation of activities may be different for immigrants, emigrants and locals.

In this chapter, we explore the differences between these proximity measures and argue that creating separate measures provides more nuance in quantifying knowledge spillovers than using a joint proximity measure.

We explore the differences in proximity measures based on the co-location of immigrants, emigrants and locals by taking the average across time (e.g. $\bar{\varphi}^{births}_{kk'} = \frac{1}{T}\sum_t \varphi^{births}_{kk',t}$). Figure S7 shows the correlation between $\bar{\varphi}^{immi}_{kk'}$, $\bar{\varphi}^{emi}_{kk'}$ and $\bar{\varphi}^{births}_{kk'}$. All proximity measures are significantly correlated with each other. For instance, $\bar{\varphi}^{emi}_{kk'}$ and $\bar{\varphi}^{births}_{kk'}$ are highly correlated with an R-squared of 0.79 (Figure S7 B).

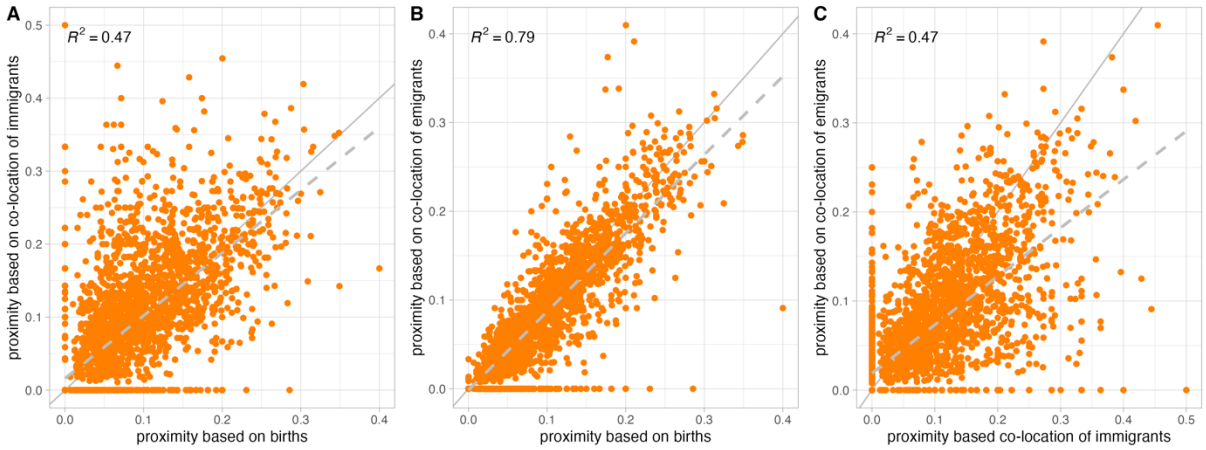



***Figure S7.*** *Correlation between the three different proximity measures*

To explore this in more detail, we take a closer look at the correlation between $\bar{\varphi}_{kk'}^{immi}$ and $\bar{\varphi}_{kk'}^{births}$. For the purposes of visualization, we restrict the sample to combinations of activities that existed in at least five centuries (Figure S8).

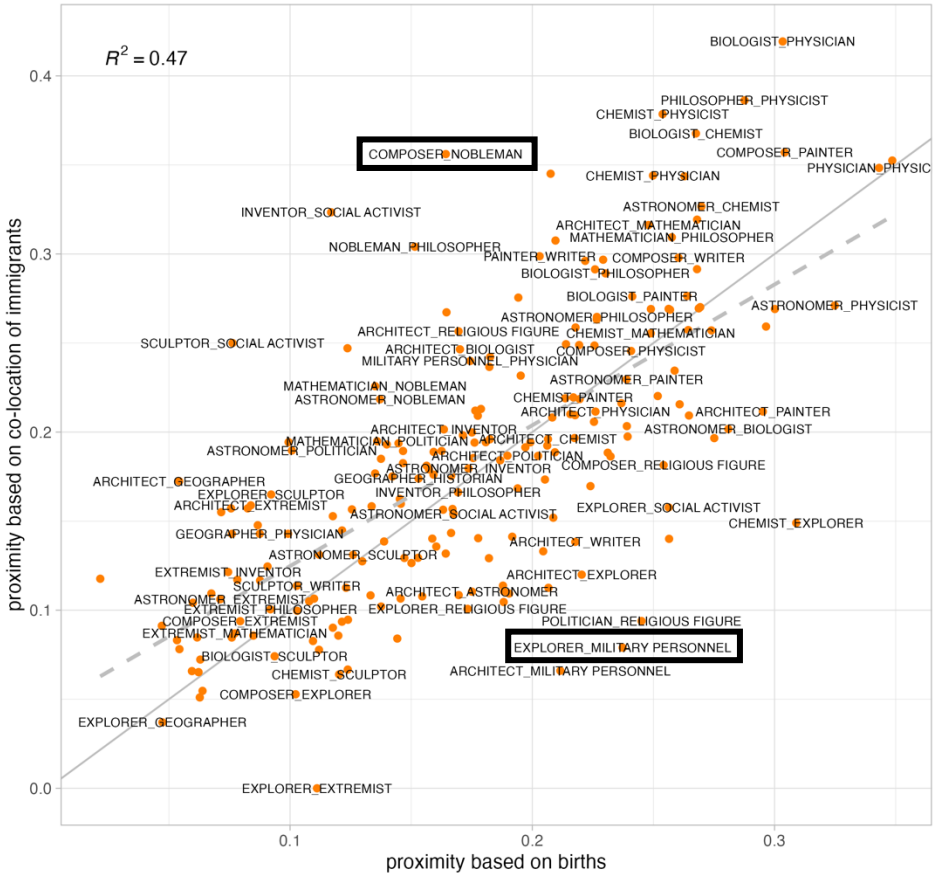



***Figure S8****. Correlation between the proximity based on births and the co-location of immigrants*

Two relevant examples (marked in the figure) emerge from this comparison.

Consider explorers and military personnel. These are highly related activities if looking at places of birth, but are distant for places of immigration. Indeed, explorers and military personnel share many required capabilities such as navigating, planning, commanding etc. (the famous Portuguese explorer Duarte Pacheco Pereira is an impersonation of this proximity). That these activities frequently co-occur in births may be explained by the fact that the institutional environment (e.g. educational structures, location at sea, defensive needs) may promote the cultivation of both these talents. If many explorers are born in a location, it, hence, makes sense that this location is related to military personnel. In contrast, the factors contributing to the immigration of explorers and military personnel seem to be less similar.

Now consider composers and noblemen. For these two activities, the proximity based on immigration patterns is higher than for the colocation of births. It makes sense that these activities are to some extent related when looking at places of birth: Noblemen are known to be patrons for the arts. Hence, noblemen born in a location will likely create institutions that promote the cultivation of the talent of composers born in this location. But it is also plausible that these activities are even more related if looking at immigration patterns. Given that we observe a disproportional migration flow of noblemen towards a certain location, we can view this location as highly related to composers, since the institutional factors attracting noblemen likely play a role in attracting and cultivating the talent of composers as well.

These examples show that separate measures of proximity provide a nuanced perspective on the relationships between activities.

Another approach is to create one joint proximity measure that does not differentiate between immigrants, emigrants and locals. To assess the robustness of our results with respect to the chosen proximity measure, we create a joint proximity measure based on the adjacency matrix $R_{ik,t}^{joint}$:

$$R_{ik,t}^{joint} = \frac{N_{ik,t}^{births} + N_{ik,t}^{deaths}}{\widehat{N}_{ik,t}^{births} + \widehat{N}_{ik,t}^{deaths}}$$

with

$$\widehat{N}_{ik,t} = \frac{\sum_k N_{ik,t} \sum_i N_{ik,t}}{\sum_{i,k} N_{ik,t}}$$

After binarizing this adjacency matrix to create $M_{ik,t}^{joint}$ we create the proximity measure $\varphi_{kk',t}^{joint}$:

$$\varphi_{kk',t}^{joint} = \frac{\sum_i M_{ik,t}^{joint} M_{ik',t}^{joint}}{max\left(\sum_i M_{ik,t}^{joint}, \sum_i M_{ik',t}^{joint}\right)}$$

Lastly, we use this joint proximity measure to create relatedness densities for immigrants, emigrants and locals:

$$\omega_{ik,t}^{immi} = \frac{\sum_{k'} M_{ik',t}^{immi} \varphi_{kk',t}^{joint}}{\sum_{k'} \varphi_{kk',t}^{joint}}$$

$$\omega_{ik,t}^{emi} = \frac{\sum_{k'} M_{ik',t}^{emi} \varphi_{kk',t}^{joint}}{\sum_{k'} \varphi_{kk',t}^{joint}}$$

$$\omega_{ik,t}^{births} = \frac{\sum_{k'} M_{ik',t}^{births} \varphi_{kk',t}^{joint}}{\sum_{k'} \varphi_{kk',t}^{joint}}$$

We use this relatedness measures in our main regression model (Eq. 11 in the main manuscript). The results (provided in the table below) are in-line with our previous findings: $\omega_{ik,t}^{immi}$ correlates positively with future entries and negatively with exits (significant at p<0.1), while the relatedness densities based on emigrants *($\omega_{ik,t}^{emi}$)* or locals *($\omega_{ik,t}^{births}$)* are insignificant. Also, the point estimates for $\omega_{ik,t}^{births}$ are lower than in our main results (Table S6 and Table S7), indicating that our finding of no robust effect for the related knowledge of locals is not driven by using separate proximity measures.

***Table S4.*** *Regression results using a joint measure of proximity.*

| Dependent Variable: | $Entry_{ik,t}$ (1) | $Exit_{ik,t}$ (2) |
|---|---|---|
| $M_{ik,t-1}^{immi}$ | 0.352*** | -0.897*** |
| | (0.036) | (0.103) |
| $M_{ik,t-1}^{emi}$ | 0.198 | -0.011 |
| | (0.297) | (0.216) |
| $\omega_{ik,t-1}^{immi}$ | 0.029* | -0.104* |
| | (0.016) | (0.063) |
| $\omega_{ik,t-1}^{emi}$ | 0.027 | -0.071 |
| | (0.030) | (0.086) |
| $\omega_{ik,t-1}^{births}$ | 0.018 | -0.010 |
| | (0.043) | (0.081) |
| $ubiquity_{k,t-1}$ | 0.008** | -0.051*** |
| | (0.003) | (0.012) |
| $\rho_{ik,t-1}^{M}$ | -0.425 | 5.904*** |
| | (0.459) | (1.342) |
| $\rho_{ik,t-1}^{\omega}$ | 0.065*** | 0.249 |
| | (0.018) | (0.173) |
| $R_{ik,t-1}$ | 0.286** | 0.003 |
| | (0.119) | (0.035) |
| FE: Broad categ.-region-period | Y | Y |
| FE: Category-period | Y | Y |
| Observations | 3944 | 1051 |
| Pseudo-$R^2$ | 0.217 | 0.238 |
| BIC | 9607.3 | 3662.8 |

Standard errors are clustered by region and period. * p < 0.1, ** p < 0.05, *** p < 0.01

# 3. Results

## 3.1. Logistic regression models explaining entries and exits and descriptive statistics

As described in the manuscript, we estimate logistic models to explain entries and exits in an activity using measures of the knowledge of immigrants and emigrants in that activity ($M_{ik,t}^{immi}$, $M_{ik,t}^{emi}$) and of the related knowledge that we can attribute to immigrants, emigrants and locals ($\omega_{ik,t}^{immi}$, $\omega_{ik,t}^{emi}$, $\omega_{ik,t}^{births}$).

To reduce endogeneity concerns because of omitted variables, we control for several other observed and unobserved factors that might influence the probability of entry or exit.

We control for an occupation's ubiquity, $\sum_i M_{ik}^{births}$ (i.e. the number of locations that are specialized in the respective occupation), since it may be easier to develop specializations in ubiquitous occupations. Also, we control for knowledge diffusion due to other channels than migration captured in spatial lags, $\rho_{ik,t-1}^{M}$ and $\rho_{ik,t-1}^{\omega}$ (see section 2.4). Lastly, our definition of entries and exits can be sensitive to borderline cases. The expected number of births may already be very close to the observed number before entering, which increases the probability of entering. Hence, we control for the ratio between the observed and expected number of births in the previous period ($R_{ik,t}^{births} = \frac{N_{ik,t}^{births}}{\tilde{N}_{ik,t}^{births}}$).

We use fixed effects to account for unobserved heterogeneity: A city may set up a university affecting both migration and future births of famous scientists. A city may also become a capital attracting politicians, journalists, or military personnel. A city may become more prosperous or increase its level of education affecting migration and future births in a field. We control for these unobserved factors by using fixed effects specific to a broad category (8 broad categories: arts, science & technology, institutions etc.; see column 1 of Table S1 for the occupation taxonomy) in a region in a century ($\gamma_{mit}$). In addition, we control for unobserved factors affecting both migration and future births that are specific to a more granular occupation category and time ($\delta_{lt}$). Index $l$ denotes one of 26 occupation categories, which distinguish, for instance, between social sciences, natural sciences and engineering within the broad category "science & technology" or music, design and film & theatre within the broad category "arts" (see column 2 in Table S1). These fixed effects capture, for instance, that the invention of motion picture technology at the end of the 19th century likely affected migration and birth patterns among film directors and actors differently than among other occupations within the same broad category of arts, such as composers or musicians.

In less restrictive specifications in section 3.4, we also add further observed control variables that have been included in the fixed effects in the main specification. That includes the number of occupations a location is specialized in (diversity, $\sum_k M_{ik}^{births}$), since the probability of entry or exit likely grows with the number of occupations already present in the respective location. Furthermore, we control for a region's population (section 1.3) at the beginning of the century ($pop_{i,t}$), because we suspect a correlation between population size and the probability of entering or exiting an activity.

Defining $Y_{ik,t} = \{Entry_{ik,t}, Exit_{ik,t}\}$, we estimate the following logistic regression model

$$\begin{aligned} P(Y_{ik,t}) = g(&\beta_1 M_{ik,t-1}^{immigrants} + \beta_2 M_{ik,t-1}^{emigrants} \\ &+ \beta_3 \omega_{ik,t-1}^{immigrants} + \beta_4 \omega_{ik,t-1}^{emigrants} \\ &+ \beta_5 \omega_{ik,t-1}^{births} + \alpha_2 ubiquity_{k,t-1} + \alpha_4 \rho_{ik,t-1}^{M} \\ &+ \alpha_5 \rho_{ik,t-1}^{\omega} + \alpha_6 R_{ik,t-1}^{births} + + \gamma_{mit} + \delta_{lt} \\ &+ \varepsilon_{ik,t}) \end{aligned} \tag{S8}$$

where $g$ denotes the logistic probability density.

Table S5 provides the summary statistics for the variables used in the regression models (based on the naïve model for the expected number of immigrants, emigrants and locals). Each entry refers to a unique combination of region, occupation and period.

The relatedness densities based on immigrants, emigrants and locals correlate with each other but show a considerable amount of variance (see Figure S9).

***Table S5**. Descriptive statistics*

| Variable | N | Mean | Std. Dev | Minimum | 25th pc. | Median | 75th pc. | Maximum |
|---|---|---|---|---|---|---|---|---|
| $\omega_{ik}^{births}$ | 43,273 | 17.83 | 10.84 | 0 | 9.65 | 15.67 | 23.87 | 100 |
| $\omega_{ik}^{immi}$ | 18,597 | 17.54 | 12.31 | 0 | 8.32 | 14.67 | 23.59 | 100 |
| $\omega_{ik}^{emi}$ | 36,985 | 17.88 | 10.59 | 0 | 9.89 | 15.85 | 23.67 | 100 |
| $M_{ik}^{births}$ | 43,814 | 0.16 | 0.36 | 0 | 0 | 0 | 0 | 1 |
| $M_{ik}^{immi}$ | 18,663 | 0.17 | 0.37 | 0 | 0 | 0 | 0 | 1 |
| $M_{ik}^{emi}$ | 37,264 | 0.15 | 0.36 | 0 | 0 | 0 | 0 | 1 |
| $Entry_{ik}$ | 15,818 | 0.17 | 0.37 | 0 | 0 | 0 | 0 | 1 |
| $Exit_{ik}$ | 3,794 | 0.64 | 0.48 | 0 | 0 | 1 | 1 | 1 |
| $N_{ik}^{births}$ | 43,814 | 0.42 | 1.60 | 0 | 0 | 0 | 0 | 64 |
| $N_{ik}^{emi}$ | 43,814 | 0.27 | 0.94 | 0 | 0 | 0 | 0 | 31 |
| $N_{ik}^{immi}$ | 43,814 | 0.27 | 1.73 | 0 | 0 | 0 | 0 | 120 |
| $diversity_i$ | 43,814 | 8.33 | 4.43 | 1 | 5 | 8 | 10 | 30 |
| $ubiquity_k$ | 43,814 | 37.45 | 34.33 | 0 | 12 | 22 | 54 | 149 |
| $\rho_{ik,t}^{\omega}$ | 43,295 | 17.76 | 5.58 | 6.87 | 13.39 | 17.06 | 21.73 | 68.14 |
| $\rho_{ik,t}^{M}$ | 43,814 | 0.16 | 0.13 | 0 | 0.05 | 0.11 | 0.24 | 0.88 |

*Note*: Each observation in the underlying dataset refers to a certain location *i*, occupation *k* and time *t*.

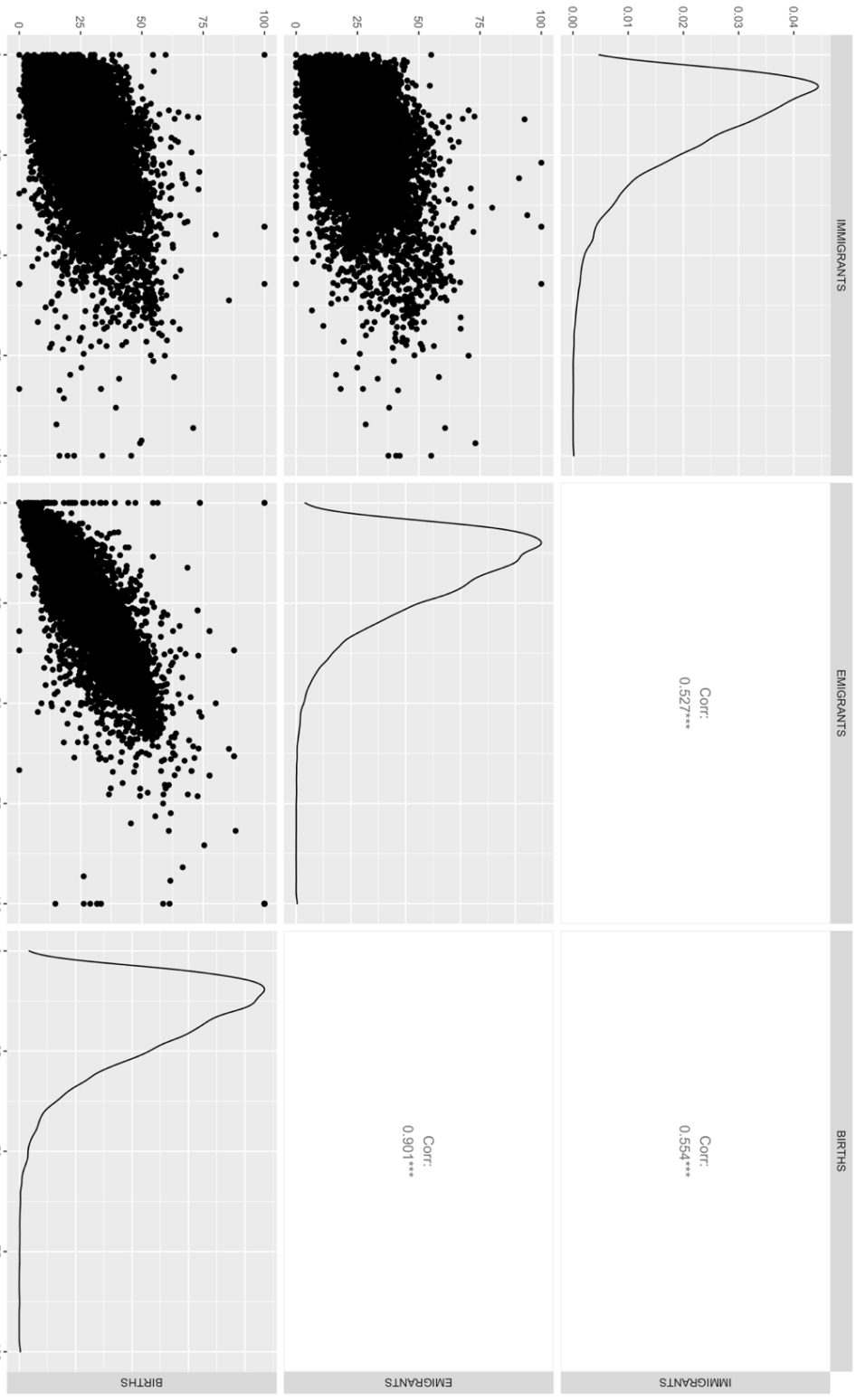



**Figure S9**. Correlations between different relatedness density metrics

## 3.2. Main regression tables explaining entries to new activities

Table S6 shows the main results for the logistic regression model explaining entries to new activities estimating the underlying expected number of immigrants, emigrants and locals with the "bins and balls" model of the Revealed Comparative Advantage. Columns 2-6 of Table S6 correspond to columns 1-5 of Table 1 in the main text.

As mentioned in the main text we find a positive correlation between entries to a specific activity and a disproportionate inflow of famous individuals with knowledge in that activity. Also, the related knowledge of immigrants correlates positively with the probability of future entries.

The control variables behave mostly as expected. We find a positive correlation between the probability of entry and the occupation's ubiquity. Thus, it is easier to enter a ubiquitous (and thus less complex) activity. Also, following the definition of entries, being closer to the threshold of a specialization ($R_{ik,t-1}$) increases the probability of entry.

***Table S6**. Main results of logistic regressions explaining the entry to new activities*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | 0.334*** | 0.303*** | 0.336*** | 0.331*** | 0.300*** |
| | | (0.080) | (0.075) | (0.086) | (0.080) | (0.076) |
| $M_{ik,t-1}^{emi}$ | | 0.115 | 0.045 | 0.106 | 0.121 | 0.018 |
| | | (0.261) | (0.278) | (0.261) | (0.255) | (0.270) |
| $\omega_{ik,t-1}^{immi}$ | | | 0.027*** | | | 0.028*** |
| | | | (0.006) | | | (0.007) |
| $\omega_{ik,t-1}^{emi}$ | | | | -0.006 | | -0.024 |
| | | | | (0.012) | | (0.019) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.011 | 0.027* |
| | | | | | (0.008) | (0.015) |
| $ubiquity_{k,t-1}$ | 0.011*** | 0.011*** | 0.010*** | 0.011*** | 0.010*** | 0.010*** |
| | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| $\rho_{ik,t-1}^{M}$ | -0.261 | -0.289 | -0.311 | -0.286 | -0.294 | -0.312 |
| | (0.562) | (0.576) | (0.559) | (0.574) | (0.583) | (0.582) |
| $\rho_{ik,t-1}^{\omega}$ | 0.096** | 0.091** | 0.075 | 0.096* | 0.081* | 0.071 |
| | (0.046) | (0.046) | (0.049) | (0.050) | (0.044) | (0.047) |
| $R_{ik,t-1}^{births}$ | 0.332*** | 0.236* | 0.295** | 0.228* | 0.246* | 0.284** |
| | (0.072) | (0.136) | (0.142) | (0.120) | (0.131) | (0.118) |
| FE: Broad categ.-region-period | Y | Y | Y | Y | Y | Y |
| FE: Category-period | Y | Y | Y | Y | Y | Y |
| Observations | 3944 | 3944 | 3944 | 3944 | 3944 | 3944 |
| Pseudo-$R^2$ | 0.211 | 0.213 | 0.214 | 0.213 | 0.213 | 0.215 |
| BIC | 9527.5 | 9537.0 | 9539.4 | 9545.0 | 9544.5 | 9553.1 |

Standard errors are clustered by period and region. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

## 3.3. Main regression tables explaining exits of activities

Table S7 shows the results of the logistic regression model explaining exits of activities estimating the underlying expected number of immigrants, emigrants and locals with the "bins and balls" model. Columns 2-6 of Table S7 correspond to columns 6-10 of Table 1 in the main text.

As mentioned in the main text we find a significantly negative correlation between exits of a specific activity and a disproportionate inflow of famous individuals with knowledge in that activity. Also, the related knowledge of immigrants correlates negatively with the probability of future exits. These coefficients are robust to other specifications and period subsets.

Considering the control variables, we find a robust negative correlation between an activity's ubiquity and the probability of exit. More ubiquitous activities exhibit a lower probability of exit. Also, we find evidence that the probability of exit grows with the presence of the same specialization in geographically close regions ($\rho^{M}_{ik,t-1}$).

***Table S7****. Main results of logistic regressions explaining the exit of activities*

| | Dependent Variable: $Exit_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M^{immi}_{ik,t-1}$ | | -0.603*** | -0.584*** | -0.591*** | -0.587*** | -0.571*** |
| | | (0.127) | (0.134) | (0.120) | (0.126) | (0.126) |
| $M^{emi}_{ik,t-1}$ | | 0.310 | 0.330 | 0.233 | 0.306 | 0.291 |
| | | (0.240) | (0.232) | (0.216) | (0.222) | (0.203) |
| $\omega^{immi}_{ik,t-1}$ | | | -0.067*** | | | -0.064*** |
| | | | (0.016) | | | (0.011) |
| $\omega^{emi}_{ik,t-1}$ | | | | -0.048 | | -0.025 |
| | | | | (0.038) | | (0.063) |
| $\omega^{births}_{ik,t-1}$ | | | | | -0.059*** | -0.034 |
| | | | | | (0.018) | (0.041) |
| $ubiquity_{k,t-1}$ | -0.053*** | -0.054*** | -0.055*** | -0.051*** | -0.051*** | -0.052*** |
| | (0.011) | (0.012) | (0.013) | (0.012) | (0.013) | (0.013) |
| $\rho^{M}_{ik,t-1}$ | 5.537*** | 5.684*** | 6.614*** | 5.127*** | 5.102*** | 5.967*** |
| | (1.273) | (1.250) | (1.453) | (1.184) | (1.180) | (1.209) |
| $\rho^{\omega}_{ik,t-1}$ | 0.118 | 0.142 | 0.145 | 0.157 | 0.152 | 0.156 |
| | (0.164) | (0.169) | (0.217) | (0.144) | (0.157) | (0.171) |
| $R^{births}_{ik,t-1}$ | 0.011 | 0.009 | -0.008 | 0.004 | 0.005 | -0.012 |
| | (0.019) | (0.019) | (0.020) | (0.016) | (0.018) | (0.018) |
| FE: Broad categ.-region-period | Y | Y | Y | Y | Y | Y |
| FE: Category-period | Y | Y | Y | Y | Y | Y |
| Observations | 1051 | 1051 | 1051 | 1051 | 1051 | 1051 |
| Pseudo-$R^2$ | 0.216 | 0.224 | 0.230 | 0.226 | 0.226 | 0.232 |
| BIC | 3616.9 | 3619.6 | 3618.0 | 3623.4 | 3623.3 | 3628.8 |

Standard errors are clustered by period and region. * p < 0.1, ** p < 0.05, *** p < 0.01

## 3.4. Robustness checks

In this section, we provide various robustness checks for our main results concerning the role of migrants in the historical geography of knowledge.

Specifically, we explore

1. potential endogeneity concerns estimating the expected number of immigrants, emigrants and locals in a negative binomial regression model (section 3.4.1)
2. different regression model specifications for both entries and exits. Because of the highly restrictive fixed effects in our main specification, the number of observations is artificially reduced. To see whether our main findings also hold for larger sample sizes, we provide results for several less restrictive fixed effects specifications (section 3.4.2)

3. the fact that distance played a more pronounced role in knowledge diffusion in earlier periods. We take this into account by interacting the measures of knowledge diffusion across space for other reasons than migration with century-dummies (section 3.4.3)
4. excluding the 20th century from the sample, since our dataset is unbalanced with respect to time (section 3.4.4)
5. a different definition of entries and exits (section 3.4.5)
6. interaction terms to investigate the role of migration in unrelated diversification (section 3.4.6)
7. heterogenous effects across different broad categories (section 3.4.7)

### 3.4.1. Estimating the expected number of immigrants, emigrants and locals

In our main specification, we defined the specialization matrices based on the concept of the Revealed Comparative Advantage. We say that a region is specialized in an activity, if the observed number of immigrants ($N_{ik,t}^{immi}$), emigrants ($N_{ik,t}^{emi}$) or locals ($N_{ik,t}^{births}$) is larger than the expected number of immigrants ($\widehat{N}_{ik,t}^{immi}$), emigrants ($\widehat{N}_{ik,t}^{emi}$) or locals ($\widehat{N}_{ik,t}^{births}$), respectively, given the size of the region and the ubiquity of the occupation (see Eq. 2 in the main text).

But based on that definition, our results shown in Table S6 and Table S7 may be subject to endogeneity. For instance, a region's local factors may change, affecting both the migration flows and the probability of giving birth to famous individuals in an activity. This could distort our estimates of whether a region is, in fact, specialized in an activity or experiences disproportionate immigration. To address these endogeneity concerns, we estimate the expected number of immigrants, emigrants and locals using not only the number of individuals in a region and the occupation's ubiquity, but also a region's specialization structure in the previous century and further unobserved factors specific to a region, activity and century.

Specifically, we estimate the following negative binomial regression models:

$$N_{ik,t}^{immi} = f(\alpha_0 + \alpha_1 N_{ik,t-1}^{immi} + \alpha_2 S_{ik,t-1}^{births} + \theta_{it} + \vartheta_{kt} + \varepsilon_{ik,t})$$
$$N_{ik,t}^{emi} = f(\beta_0 + \beta_1 N_{ik,t-1}^{emi} + \beta_2 S_{ik,t-1}^{births} + \theta_{it} + \vartheta_{kt} + \varepsilon_{ik,t}) \quad , \quad \text{(S9)}$$
$$N_{ik,t}^{births} = f(\gamma_0 + \gamma_1 N_{ik,t-1}^{births} + \gamma_2 S_{ik,t-1}^{births} + \theta_{it} + \vartheta_{kt} + \varepsilon_{ik,t})$$

where $S_{ik,t-1}^{births} = \frac{N_{ik}^{births} / \sum_k N_{ik}^{births}}{\sum_i N_{ik}^{births} / \sum_{i,k} N_{ik}^{births}}$, while $\theta_{it}$ and $\vartheta_{kt}$ denote fixed-effects accounting for unobserved factors specific to a region in a specific century and to an occupation in a specific century, respectively. Table S8 shows the results.

We use the fitted values of these regression models as the expected values in creating the specialization matrices in Eq. S2 ($\widehat{N}_{ik,t}^{immi}$, $\widehat{N}_{ik,t}^{emi}$, $\widehat{N}_{ik,t}^{births}$).

Using these adjacency matrices, we calculate the proximities between activities as well as the relatedness densities for immigrants, emigrants and locals. Figure S10 shows that the original measures of related knowledge based on the more naïve definition of the expected numbers and the relatedness densities based on the expected numbers retrieved from Eq. S9 are highly correlated with an $R^2$ of 0.9 to 0.95.

We then use these new measures of the knowledge of immigrants, emigrants and locals in the logistic regression models described in Eq. S8. The results for both entries (Table S9) and exits (Table S10) remain qualitatively unchanged for the knowledge of immigrants.

***Table S8***. *Negative binomial regression models to estimate the expected number of immigrants, emigrants and locals.*

| | $N_{ik,t}^{immi}$ | $N_{ik,t}^{emi}$ | $N_{ik,t}^{births}$ |
|---|---|---|---|
| $N_{ik,t-1}^{immi}$ | 0.033** | | |
| | (0.013) | | |
| $N_{ik,t-1}^{emi}$ | | 0.038*** | |
| | | (0.011) | |
| $N_{ik,t-1}^{births}$ | | | 0.033*** |
| | | | (0.009) |
| $S_{ik,t-1}^{births}$ | 0.045*** | 0.038*** | 0.040*** |
| | (0.006) | (0.005) | (0.005) |
| Overdispersion parameter | 2.033*** | 3.453*** | 2.369*** |
| | (0.163) | (0.294) | (0.138) |
| FE: period-region | X | X | X |
| FE: period-occupation | X | X | X |
| Num.Obs. | 39131 | 43651 | 43755 |
| Pseudo-$R^2$ | 0.341 | 0.292 | 0.282 |
| AIC | 30775.6 | 39170.4 | 50334.3 |
| BIC | 40173.4 | 49573.8 | 60775.3 |

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

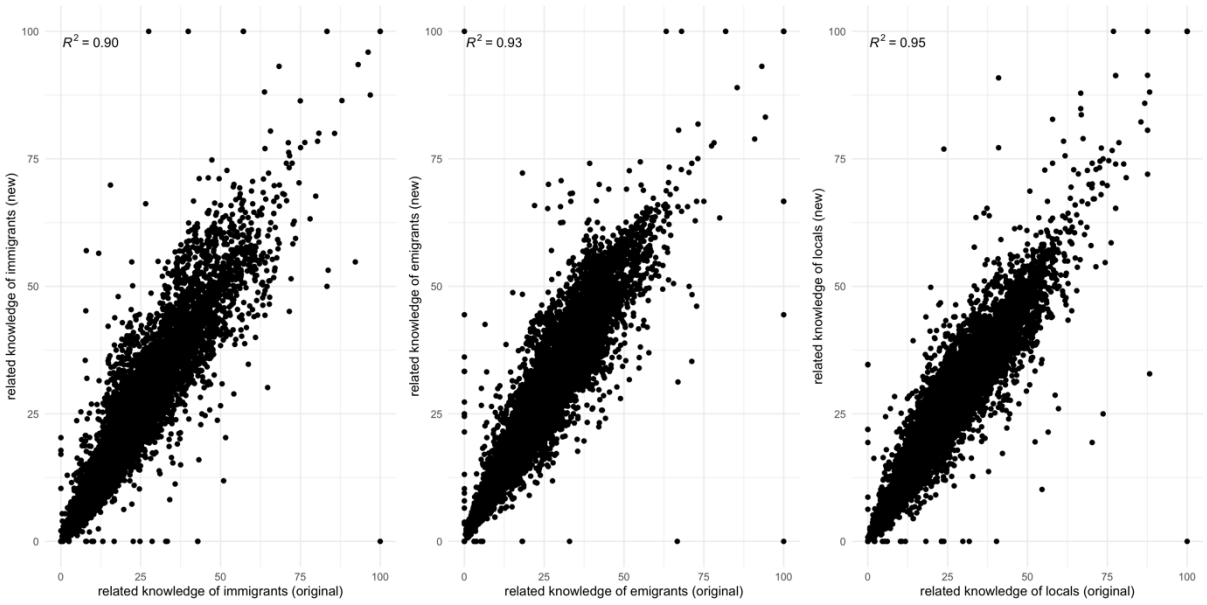



***Figure S10***. *Correlation between the relatedness densities based on the definition of expected number as in the Revealed Comparative Advantage (original) or the negative binomial regression (new).*

***Table S9***. *Logistic regression model explaining entries to new activities using the expected numbers of the model in Eq. S11*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | 0.295** | 0.254** | 0.295** | 0.293** | 0.251*** |
| | | (0.140) | (0.138) | (0.140) | (0.138) | (0.058) |
| $M_{ik,t-1}^{emi}$ | | 0.347 | 0.355 | 0.347 | 0.334 | 0.316 |
| | | (0.372) | (0.364) | (0.374) | (0.372) | (0.325) |
| $\omega_{ik,t-1}^{immi}$ | | | 0.027** | | | 0.026*** |
| | | | (0.013) | | | (0.008) |
| $\omega_{ik,t-1}^{emi}$ | | | | 0.0003 | | -0.026 |
| | | | | (0.014) | | (0.019) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.028** | 0.044** |
| | | | | | (0.015) | (0.016) |
| $ubiquity_{k,t-1}$ | 0.011 | 0.011 | 0.010 | 0.011 | 0.010 | 0.010*** |
| | (0.008) | (0.007) | (0.007) | (0.007) | (0.007) | (0.003) |
| $\rho_{ik,t-1}^{M}$ | -0.261 | -0.289 | -0.311 | -0.286 | -0.294 | -0.312 |
| | (0.916) | (0.932) | (0.948) | (0.934) | (0.932) | (0.582) |
| $\rho_{ik,t-1}^{\omega}$ | 0.096 | 0.091 | 0.075 | 0.096 | 0.081 | 0.071 |
| | (0.059) | (0.059) | (0.063) | (0.060) | (0.056) | (0.047) |
| $R_{ik,t-1}^{births}$ | 0.332 | 0.236 | 0.295 | 0.228 | 0.246 | 0.284** |
| | (0.262) | (0.271) | (0.271) | (0.274) | (0.273) | (0.118) |
| FE: Broad categ.-region-period | Y | Y | Y | Y | Y | Y |
| FE: Category-period | Y | Y | Y | Y | Y | Y |
| Observations | 4049 | 4049 | 4049 | 4049 | 4049 | 4049 |
| Pseudo-$R^2$ | 0.224 | 0.225 | 0.227 | 0.225 | 0.226 | 0.228 |
| BIC | 9796.7 | 9806.5 | 9807.7 | 9814.8 | 9810.7 | 9818.0 |

Standard errors are clustered by period and region. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

***Table S10**. Logistic regression model explaining exits from existing areas of specialization using the expected numbers of the model in Eq. S11*

| | Dependent Variable: $Exit_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | -0.411* | -0.383* | -0.409* | -0.411* | -0.388** |
| | | (0.226) | (0.218) | (0.225) | (0.226) | (0.159) |
| $M_{ik,t-1}^{emi}$ | | 0.851** | 0.864** | 0.822** | 0.828** | 0.884*** |
| | | (0.337) | (0.339) | (0.328) | (0.339) | (0.206) |
| $\omega_{ik,t-1}^{immi}$ | | | -0.042 | | | -0.040* |
| | | | (0.028) | | | (0.022) |
| $\omega_{ik,t-1}^{emi}$ | | | | -0.018 | | 0.028 |
| | | | | (0.047) | | (0.047) |
| $\omega_{ik,t-1}^{births}$ | | | | | -0.063 | -0.080** |
| | | | | | (0.041) | (0.041) |
| $ubiquity_{k,t-1}$ | -0.052*** | -0.053*** | -0.054*** | -0.052*** | -0.050*** | -0.053*** |
| | (0.012) | (0.012) | (0.012) | (0.012) | (0.012) | (0.009) |
| $\rho_{ik,t-1}^{M}$ | 4.793** | 4.550** | 4.886** | 4.450* | 4.207* | 4.621* |
| | (2.310) | (2.254) | (2.277) | (2.331) | (2.236) | (2.139) |
| $\rho_{ik,t-1}^{\omega}$ | 0.209 | 0.241* | 0.262* | 0.243* | 0.247* | 0.265** |
| | (0.128) | (0.133) | (0.133) | (0.132) | (0.127) | (0.072) |
| $R_{ik,t-1}^{births}$ | 0.028 | 0.022 | 0.012 | 0.022 | 0.020 | 0.010 |
| | (0.039) | (0.038) | (0.035) | (0.037) | (0.036) | (0.024) |
| FE: Broad categ.-region-period | Y | Y | Y | Y | Y | Y |
| FE: Category-period | Y | Y | Y | Y | Y | Y |
| Observations | 1084 | 1084 | 1084 | 1084 | 1084 | 1084 |
| Pseudo-$R^2$ | 0.195 | 0.205 | 0.209 | 0.206 | 0.208 | 0.212 |
| BIC | 3804.0 | 3802.9 | 3805.0 | 3809.5 | 3805.3 | 3814.1 |

Standard errors are clustered by period and region. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

### 3.4.2. Model specifications

To control for unobserved factors that have an influence on both the probability of entering or exiting an activity and migration flows, we use fixed effects. For example, period fixed effects account for any unobserved heterogeneity that is time-specific, but independent of the location and the activity. Thus, this controls for the effect that entering a new activity may have become more easier over time, because of economic development and urbanization.

In the main results, we employ a highly restrictive fixed effects structure accounting for unobserved factors specific to a region, century and broad category as well as unobserved factors specific to an occupation category and century (see Eq. S8, Table S6 and Table S7).

Although this specification addresses endogeneity issues thoroughly, this highly restrictive fixed effects specification (more than 700 fixed effect coefficients) comes at a cost, too. That is, our sample size is artificially reduced, since observations with no changes in the dependent variable within a fixed effect category are removed. In this section, we provide several less restrictive fixed effects specifications to see how our results change.

Furthermore, a less restrictive fixed effects specification allows for including more control variables that previously were included in the fixed effects. That includes the number of occupations a location is specialized in (diversity, $\sum_k M_{ik}^{births}$), since the probability of entry or

exit likely grows with the number of occupations already present in the respective location. Furthermore, we control for a region's population (section 1.3) at the beginning of the century ($pop_{i,t}$), because we suspect a correlation between population size and the probability of entering or exiting an activity.

In the following, we control either for time-location and occupation category fixed effects (Table S11 and Table S15 for entries and exits, respectively), for time, location and occupation category fixed effects (Table S12 and Table S16 for entries and exits, respectively), for time and location fixed effects (Table S13 and Table S17 for entries and exits, respectively) or only for time fixed effects (Table S14 and Table S18 for entries and exits, respectively). Despite the different fixed effects, the main results for $M_{ik,t-1}^{immi}$ and $\omega_{ik,t-1}^{immi}$ remain unchanged.

***Table S11***. *Logistic regression model explaining entries, accounting for period-region and occupation category fixed effects*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | 0.389*** | 0.339*** | 0.392*** | 0.379*** | 0.326*** |
| | | (0.095) | (0.084) | (0.095) | (0.096) | (0.083) |
| $M_{ik,t-1}^{emi}$ | | 0.033 | 0.011 | 0.026 | 0.031 | -0.017 |
| | | (0.357) | (0.369) | (0.362) | (0.353) | (0.373) |
| $\omega_{ik,t-1}^{immi}$ | | | 0.019** | | | 0.020** |
| | | | (0.009) | | | (0.009) |
| $\omega_{ik,t-1}^{emi}$ | | | | -0.007 | | -0.024** |
| | | | | (0.009) | | (0.010) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.013* | 0.029*** |
| | | | | | (0.007) | (0.007) |
| $ubiquity_{k,t-1}$ | 0.007 | 0.007 | 0.006 | 0.007 | 0.007 | 0.007 |
| | (0.005) | (0.005) | (0.004) | (0.004) | (0.005) | (0.004) |
| $\rho_{ik,t-1}^{M}$ | 0.652 | 0.588 | 0.604 | 0.592 | 0.597 | 0.620 |
| | (0.721) | (0.729) | (0.716) | (0.731) | (0.730) | (0.719) |
| $\rho_{ik,t-1}^{\omega}$ | 0.034 | 0.034 | 0.025 | 0.039 | 0.021 | 0.015 |
| | (0.031) | (0.031) | (0.031) | (0.031) | (0.031) | (0.032) |
| $R_{ik,t-1}^{births}$ | 0.525** | 0.463* | 0.493* | 0.458* | 0.471* | 0.496* |
| | (0.249) | (0.265) | (0.273) | (0.268) | (0.258) | (0.272) |
| FE: period-region | Y | Y | Y | Y | Y | Y |
| FE: occu. category | Y | Y | Y | Y | Y | Y |
| Observations | 6165 | 6165 | 6165 | 6165 | 6165 | 6165 |
| Pseudo-$R^2$ | 0.131 | 0.134 | 0.136 | 0.135 | 0.135 | 0.137 |
| AIC | 5782.5 | 5768.7 | 5762.6 | 5770.1 | 5768.3 | 5758.7 |
| BIC | 7437.3 | 7437.0 | 7437.5 | 7445.0 | 7443.3 | 7447.1 |

Standard errors are clustered by region and period. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

***Table S12***. *Logistic regression model explaining entries, accounting for period-, location- and occupation category-fixed effects*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | 0.380*** | 0.335*** | 0.381*** | 0.375*** | 0.330*** |
| | | (0.073) | (0.066) | (0.074) | (0.074) | (0.066) |
| $M_{ik,t-1}^{emi}$ | | 0.012 | -0.027 | 0.015 | 0.010 | -0.017 |
| | | (0.235) | (0.228) | (0.231) | (0.234) | (0.224) |
| $\omega_{ik,t-1}^{immi}$ | | | 0.011*** | | | 0.012*** |
| | | | (0.003) | | | (0.002) |
| $\omega_{ik,t-1}^{emi}$ | | | | -0.001 | | -0.008*** |
| | | | | (0.004) | | (0.002) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.006 | 0.011** |
| | | | | | (0.006) | (0.006) |
| $diversity_{i,t-1}$ | -0.008 | -0.007 | -0.009 | -0.005 | -0.021 | -0.022 |
| | (0.020) | (0.020) | (0.020) | (0.017) | (0.020) | (0.022) |
| $ubiquity_{k,t-1}$ | 0.008*** | 0.007*** | 0.007*** | 0.007*** | 0.007*** | 0.007*** |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| $\rho_{ik,t-1}^{M}$ | 0.622*** | 0.562*** | 0.590*** | 0.562*** | 0.568*** | 0.598*** |
| | (0.207) | (0.159) | (0.122) | (0.161) | (0.175) | (0.155) |
| $\rho_{ik,t-1}^{\omega}$ | 0.029 | 0.029 | 0.021 | 0.030 | 0.023 | 0.016 |
| | (0.025) | (0.024) | (0.024) | (0.024) | (0.024) | (0.024) |
| $R_{ik,t-1}^{births}$ | 0.533*** | 0.478** | 0.493*** | 0.477** | 0.483*** | 0.499*** |
| | (0.124) | (0.187) | (0.178) | (0.185) | (0.182) | (0.178) |
| log $(pop_{i,t})$ | 0.368*** | 0.351*** | 0.249*** | 0.352*** | 0.349*** | 0.249*** |
| | (0.067) | (0.074) | (0.089) | (0.074) | (0.069) | (0.083) |
| FE: period | Y | Y | Y | Y | Y | Y |
| FE: region | Y | Y | Y | Y | Y | Y |
| FE: occu. category | Y | Y | Y | Y | Y | Y |
| Observations | 6180 | 6180 | 6180 | 6180 | 6180 | 6180 |
| Pseudo-$R^2$ | 0.121 | 0.124 | 0.124 | 0.124 | 0.124 | 0.125 |
| AIC | 5694.6 | 5680.9 | 5677.5 | 5682.8 | 5682.1 | 5679.4 |
| BIC | 6811.6 | 6811.4 | 6814.7 | 6820.0 | 6819.3 | 6830.0 |

Standard errors are clustered by period and region. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

***Table S13***. *Logistic regression model explaining entries, accounting for period- and location-fixed effects*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | 0.367*** | 0.324*** | 0.368*** | 0.361*** | 0.317*** |
| | | (0.086) | (0.076) | (0.087) | (0.087) | (0.076) |
| $M_{ik,t-1}^{emi}$ | | 0.207 | 0.174 | 0.208 | 0.204 | 0.182 |
| | | (0.223) | (0.220) | (0.219) | (0.221) | (0.215) |
| $\omega_{ik,t-1}^{immi}$ | | | 0.010*** | | | 0.011*** |
| | | | (0.003) | | | (0.003) |
| $\omega_{ik,t-1}^{emi}$ | | | | -0.001 | | -0.008*** |
| | | | | (0.003) | | (0.002) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.008 | 0.013* |
| | | | | | (0.007) | (0.007) |
| $diversity_{i,t-1}$ | -0.012 | -0.011 | -0.013 | -0.010 | -0.028 | -0.030 |
| | (0.019) | (0.019) | (0.019) | (0.015) | (0.019) | (0.020) |
| $ubiquity_{k,t-1}$ | 0.009*** | 0.009*** | 0.009*** | 0.009*** | 0.009*** | 0.009*** |
| | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| $\rho_{ik,t-1}^{M}$ | 0.290 | 0.245* | 0.255* | 0.245* | 0.259* | 0.273 |
| | (0.185) | (0.128) | (0.152) | (0.128) | (0.154) | (0.186) |
| $\rho_{ik,t-1}^{\omega}$ | 0.029 | 0.026 | 0.020 | 0.027 | 0.020 | 0.013 |
| | (0.045) | (0.043) | (0.044) | (0.043) | (0.040) | (0.041) |
| $R_{ik,t-1}^{births}$ | 0.298*** | 0.161 | 0.171 | 0.160 | 0.170 | 0.184 |
| | (0.114) | (0.181) | (0.173) | (0.180) | (0.171) | (0.170) |
| log $(pop_{i,t})$ | 0.363*** | 0.350*** | 0.257*** | 0.352*** | 0.349*** | 0.258*** |
| | (0.055) | (0.061) | (0.074) | (0.061) | (0.057) | (0.068) |
| FE: period | Y | Y | Y | Y | Y | Y |
| FE: region | Y | Y | Y | Y | Y | Y |
| Observations | 6180 | 6180 | 6180 | 6180 | 6180 | 6180 |
| Pseudo-$R^2$ | 0.092 | 0.095 | 0.096 | 0.095 | 0.095 | 0.096 |
| AIC | 5819.7 | 5806.0 | 5803.1 | 5808.0 | 5806.7 | 5804.3 |
| BIC | 6775.3 | 6775.0 | 6778.8 | 6783.7 | 6782.4 | 6793.5 |

Standard errors are clustered by region and period. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

***Table S14**. Logistic regression model explaining entries, accounting for period-fixed effects*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | 0.406*** | 0.327*** | 0.404*** | 0.403*** | 0.322*** |
| | | (0.072) | (0.076) | (0.072) | (0.073) | (0.077) |
| $M_{ik,t-1}^{emi}$ | | 0.195 | 0.153 | 0.189 | 0.194 | 0.147 |
| | | (0.175) | (0.171) | (0.176) | (0.173) | (0.169) |
| $\omega_{ik,t-1}^{immi}$ | | | 0.011*** | | | 0.011*** |
| | | | (0.002) | | | (0.003) |
| $\omega_{ik,t-1}^{emi}$ | | | | 0.004 | | 0.003 |
| | | | | (0.002) | | (0.003) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.005 | 0.004 |
| | | | | | (0.005) | (0.006) |
| $diversity_{i,t-1}$ | 0.041*** | 0.037*** | 0.025*** | 0.030*** | 0.028*** | 0.012 |
| | (0.004) | (0.005) | (0.007) | (0.006) | (0.011) | (0.015) |
| $ubiquity_{k,t-1}$ | 0.008*** | 0.008*** | 0.008*** | 0.007*** | 0.007*** | 0.007*** |
| | (0.003) | (0.002) | (0.002) | (0.003) | (0.003) | (0.003) |
| $\rho_{ik,t-1}^{M}$ | 0.235 | 0.186 | 0.208 | 0.188 | 0.192 | 0.215 |
| | (0.337) | (0.323) | (0.329) | (0.325) | (0.328) | (0.336) |
| $\rho_{ik,t-1}^{\omega}$ | 0.040 | 0.037 | 0.029 | 0.035 | 0.034 | 0.025 |
| | (0.039) | (0.038) | (0.035) | (0.037) | (0.035) | (0.032) |
| $R_{ik,t-1}^{births}$ | 0.366*** | 0.224 | 0.219 | 0.222 | 0.230 | 0.223 |
| | (0.111) | (0.163) | (0.157) | (0.160) | (0.156) | (0.148) |
| log $(pop_{i,t})$ | 0.164*** | 0.158** | 0.130* | 0.160** | 0.161** | 0.134* |
| | (0.062) | (0.067) | (0.070) | (0.067) | (0.069) | (0.071) |
| FE: period | Y | Y | Y | Y | Y | Y |
| Observations | 6180 | 6180 | 6180 | 6180 | 6180 | 6180 |
| Pseudo-$R^2$ | 0.070 | 0.073 | 0.075 | 0.073 | 0.073 | 0.075 |
| AIC | 5697.9 | 5679.1 | 5670.1 | 5680.3 | 5680.5 | 5672.9 |
| BIC | 5778.6 | 5773.3 | 5771.1 | 5781.3 | 5781.4 | 5787.3 |

Standard errors are clustered by region and period. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S15**. Logistic regression model explaining exits, accounting for period-region and occupation category fixed effects*

| | Dependent Variable: $Exit_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | -0.565*** | -0.523*** | -0.563*** | -0.553*** | -0.514*** |
| | | (0.155) | (0.148) | (0.155) | (0.157) | (0.150) |
| $M_{ik,t-1}^{emi}$ | | 0.142 | 0.148 | 0.129 | 0.151 | 0.157 |
| | | (0.168) | (0.173) | (0.169) | (0.169) | (0.174) |
| $\omega_{ik,t-1}^{immi}$ | | | -0.037*** | | | -0.036*** |
| | | | (0.014) | | | (0.014) |
| $\omega_{ik,t-1}^{emi}$ | | | | -0.016 | | 0.000 |
| | | | | (0.012) | | (0.017) |
| $\omega_{ik,t-1}^{births}$ | | | | | -0.040*** | -0.038* |
| | | | | | (0.015) | (0.021) |
| $ubiquity_{k,t-1}$ | -0.027*** | -0.026*** | -0.026*** | -0.025*** | -0.024*** | -0.024*** |
| | (0.007) | (0.006) | (0.006) | (0.007) | (0.007) | (0.007) |
| $\rho_{ik,t-1}^{M}$ | 1.371 | 1.427* | 1.685** | 1.297 | 1.122 | 1.399* |
| | (0.915) | (0.848) | (0.807) | (0.801) | (0.851) | (0.796) |
| $\rho_{ik,t-1}^{\omega}$ | 0.033 | 0.040 | 0.037 | 0.037 | 0.039 | 0.037 |
| | (0.049) | (0.052) | (0.057) | (0.051) | (0.052) | (0.056) |
| $R_{ik,t-1}^{births}$ | -0.021 | -0.019 | -0.026 | -0.020 | -0.022 | -0.029 |
| | (0.022) | (0.024) | (0.026) | (0.024) | (0.024) | (0.026) |
| FE: period-region | Y | Y | Y | Y | Y | Y |
| FE: occu. category | Y | Y | Y | Y | Y | Y |
| Observations | 1989 | 1989 | 1989 | 1989 | 1989 | 1989 |
| Pseudo-$R^2$ | 0.159 | 0.168 | 0.172 | 0.168 | 0.170 | 0.173 |
| AIC | 2665.7 | 2646.5 | 2637.9 | 2647.0 | 2643.1 | 2637.3 |
| BIC | 3969.4 | 3961.4 | 3958.5 | 3967.5 | 3963.6 | 3969.0 |

Standard errors are clustered by region and period. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S16**. Logistic regression model explaining exits, accounting for period-, location- and occupation category-fixed effects*

| | Dependent Variable: $Exit_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | -0.509*** | -0.464*** | -0.509*** | -0.510*** | -0.461*** |
| | | (0.051) | (0.064) | (0.061) | (0.051) | (0.072) |
| $M_{ik,t-1}^{emi}$ | | 0.131 | 0.123 | 0.125 | 0.131 | 0.114 |
| | | (0.177) | (0.180) | (0.185) | (0.177) | (0.190) |
| $\omega_{ik,t-1}^{immi}$ | | | -0.014* | | | -0.014** |
| | | | (0.008) | | | (0.007) |
| $\omega_{ik,t-1}^{emi}$ | | | | 0.005 | | 0.008 |
| | | | | (0.015) | | (0.019) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.001 | -0.002 |
| | | | | | (0.006) | (0.012) |
| $diversity_{i,t-1}$ | -0.042 | -0.044 | -0.039 | -0.051 | -0.046 | -0.045 |
| | (0.029) | (0.032) | (0.033) | (0.044) | (0.036) | (0.039) |
| $ubiquity_{k,t-1}$ | -0.027*** | -0.026*** | -0.025*** | -0.026*** | -0.026*** | -0.026*** |
| | (0.006) | (0.006) | (0.006) | (0.005) | (0.006) | (0.005) |
| $\rho_{ik,t-1}^{M}$ | 1.399* | 1.405 | 1.417 | 1.459* | 1.417 | 1.473* |
| | (0.836) | (0.873) | (0.882) | (0.844) | (0.896) | (0.891) |
| $\rho_{ik,t-1}^{\omega}$ | 0.038 | 0.042 | 0.040 | 0.043 | 0.043 | 0.042 |
| | (0.050) | (0.053) | (0.052) | (0.050) | (0.054) | (0.051) |
| $R_{ik,t-1}^{births}$ | -0.023*** | -0.022*** | -0.026*** | -0.021*** | -0.022*** | -0.025*** |
| | (0.006) | (0.007) | (0.008) | (0.006) | (0.007) | (0.007) |
| log $(pop_{i,t})$ | -0.372 | -0.340 | -0.265 | -0.357 | -0.343 | -0.283 |
| | (0.277) | (0.272) | (0.272) | (0.276) | (0.278) | (0.280) |
| FE: period | Y | Y | Y | Y | Y | Y |
| FE: region | Y | Y | Y | Y | Y | Y |
| FE: occu. category | Y | Y | Y | Y | Y | Y |
| Observations | 2017 | 2017 | 2017 | 2017 | 2017 | 2017 |
| Pseudo-$R^2$ | 0.138 | 0.145 | 0.147 | 0.146 | 0.145 | 0.147 |
| AIC | 2624.5 | 2607.6 | 2605.7 | 2609.2 | 2609.6 | 2608.9 |
| BIC | 3533.3 | 3527.5 | 3531.2 | 3534.7 | 3535.1 | 3545.6 |

Standard errors are clustered by region and period. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S17**. Logistic regression model explaining exits, accounting for period- and location-fixed effects*

| | Dependent Variable: $Exit_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | -0.520*** | -0.464*** | -0.520*** | -0.521*** | -0.460*** |
| | | (0.034) | (0.033) | (0.041) | (0.034) | (0.042) |
| $M_{ik,t-1}^{emi}$ | | 0.060 | 0.053 | 0.052 | 0.060 | 0.042 |
| | | (0.182) | (0.185) | (0.191) | (0.182) | (0.195) |
| $\omega_{ik,t-1}^{immi}$ | | | -0.016*** | | | -0.017*** |
| | | | (0.006) | | | (0.006) |
| $\omega_{ik,t-1}^{emi}$ | | | | 0.007 | | 0.010 |
| | | | | (0.012) | | (0.012) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.003 | -0.002 |
| | | | | | (0.006) | (0.007) |
| $diversity_{i,t-1}$ | -0.036 | -0.037 | -0.031 | -0.047 | -0.042 | -0.041 |
| | (0.027) | (0.028) | (0.027) | (0.039) | (0.029) | (0.030) |
| $ubiquity_{k,t-1}$ | -0.020*** | -0.019*** | -0.018*** | -0.019*** | -0.019*** | -0.019*** |
| | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) |
| $\rho_{ik,t-1}^{M}$ | 1.277 | 1.305 | 1.317 | 1.384 | 1.332 | 1.404 |
| | (0.891) | (0.956) | (1.007) | (0.923) | (0.976) | (1.009) |
| $\rho_{ik,t-1}^{\omega}$ | 0.027 | 0.032 | 0.031 | 0.033 | 0.032 | 0.033 |
| | (0.045) | (0.049) | (0.049) | (0.047) | (0.049) | (0.048) |
| $R_{ik,t-1}^{births}$ | -0.012** | -0.011* | -0.015*** | -0.010* | -0.011 | -0.014** |
| | (0.005) | (0.006) | (0.005) | (0.006) | (0.007) | (0.006) |
| log $(pop_{i,t})$ | -0.316 | -0.284 | -0.196 | -0.308 | -0.291 | -0.221 |
| | (0.257) | (0.251) | (0.249) | (0.252) | (0.256) | (0.254) |
| FE: period | Y | Y | Y | Y | Y | Y |
| FE: region | Y | Y | Y | Y | Y | Y |
| Observations | 2023 | 2023 | 2023 | 2023 | 2023 | 2023 |
| Pseudo-$R^2$ | 0.102 | 0.111 | 0.113 | 0.111 | 0.111 | 0.114 |
| AIC | 2677.9 | 2659.0 | 2655.1 | 2660.1 | 2661.0 | 2657.5 |
| BIC | 3458.0 | 3450.4 | 3452.0 | 3457.1 | 3457.9 | 3465.7 |

Standard errors are clustered by region and period. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S18**. Logistic regression model explaining exits, accounting for period-fixed effects*

| | Dependent Variable: $Exit_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | | -0.549*** | -0.469*** | -0.541*** | -0.550*** | -0.464*** |
| | | (0.043) | (0.047) | (0.047) | (0.042) | (0.054) |
| $M_{ik,t-1}^{emi}$ | | 0.159 | 0.131 | 0.121 | 0.158 | 0.097 |
| | | (0.181) | (0.177) | (0.189) | (0.184) | (0.195) |
| $\omega_{ik,t-1}^{immi}$ | | | -0.012*** | | | -0.012*** |
| | | | (0.003) | | | (0.003) |
| $\omega_{ik,t-1}^{emi}$ | | | | 0.011 | | 0.009 |
| | | | | (0.008) | | (0.009) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.011 | 0.004 |
| | | | | | (0.008) | (0.008) |
| $diversity_{i,t-1}$ | -0.043*** | -0.039*** | -0.024*** | -0.058*** | -0.063*** | -0.051** |
| | (0.006) | (0.004) | (0.006) | (0.017) | (0.017) | (0.020) |
| $ubiquity_{k,t-1}$ | -0.014*** | -0.013*** | -0.013*** | -0.014*** | -0.014*** | -0.014*** |
| | (0.003) | (0.004) | (0.004) | (0.003) | (0.004) | (0.003) |
| $\rho_{ik,t-1}^{M}$ | 1.035 | 1.056 | 1.022 | 1.188 | 1.176 | 1.183 |
| | (0.782) | (0.872) | (0.904) | (0.802) | (0.838) | (0.837) |
| $\rho_{ik,t-1}^{\omega}$ | -0.013 | -0.006 | -0.001 | -0.006 | -0.005 | -0.001 |
| | (0.025) | (0.025) | (0.027) | (0.025) | (0.025) | (0.027) |
| $R_{ik,t-1}^{births}$ | -0.007 | -0.008 | -0.010** | -0.006 | -0.007 | -0.008 |
| | (0.006) | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) |
| log $(pop_{i,t})$ | -0.265*** | -0.247*** | -0.218*** | -0.246*** | -0.241*** | -0.215*** |
| | (0.019) | (0.022) | (0.031) | (0.023) | (0.019) | (0.031) |
| FE: period | Y | Y | Y | Y | Y | Y |
| Observations | 2045 | 2045 | 2045 | 2045 | 2045 | 2045 |
| Pseudo-$R^2$ | 0.059 | 0.070 | 0.073 | 0.072 | 0.071 | 0.074 |
| AIC | 2558.5 | 2531.7 | 2526.2 | 2530.7 | 2532.3 | 2527.1 |
| BIC | 2626.0 | 2610.5 | 2610.6 | 2615.1 | 2616.6 | 2622.7 |

Standard errors are clustered by region and period. * p < 0.1, ** p < 0.05, *** p < 0.01

### 3.4.3. Century-specific distance measures

Furthermore, we acknowledge that distances across space, if looking at such long periods, are not constant over time, but decrease due to improvements in the infrastructure or technological progress. Hence, we interact our measures of spatial proximity, $\rho_{ik,t-1}^{M}$ and $\rho_{ik,t-1}^{\omega}$ (see section 2.4), with dummies indicating the different centuries to alleviate concerns that our results are subject to omitted variable bias. Table S19 and Table S20 show that the results remain unchanged for both entries and exits, respectively.

***Table S19***. *Logistic regression model explaining entries to new activities, interacting measures of spatial proximity with period fixed-effects.*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immigrants}$ | | 0.322*** | 0.289*** | 0.324** | 0.318*** | 0.284** |
| | | (0.120) | (0.108) | (0.127) | (0.120) | (0.121) |
| $M_{ik,t-1}^{emigrants}$ | | 0.163 | 0.083 | 0.144 | 0.171 | 0.044 |
| | | (0.469) | (0.464) | (0.503) | (0.493) | (0.570) |
| $\omega_{ik,t-1}^{immigrants}$ | | | 0.031*** | | | 0.032** |
| | | | (0.010) | | | (0.015) |
| $\omega_{ik,t-1}^{emigrants}$ | | | | -0.010 | | -0.030 |
| | | | | (0.019) | | (0.029) |
| $\omega_{ik,t-1}^{births}$ | | | | | 0.010 | 0.029 |
| | | | | | (0.021) | (0.030) |
| $ubiquity_{k,t-1}$ | 0.003 | 0.003 | 0.001 | 0.003 | 0.002 | 0.001 |
| | (0.004) | (0.004) | (0.004) | (0.004) | (0.003) | (0.005) |
| $\rho_{ik,t-1}^{M}$ | 0.696** | 0.744** | 0.751** | 0.748** | 0.772** | 0.848*** |
| | (0.314) | (0.338) | (0.326) | (0.349) | (0.334) | (0.308) |
| $\rho_{ik,t-1}^{\omega}$ | 0.146*** | 0.142*** | 0.131*** | 0.152*** | 0.131*** | 0.128*** |
| | (0.051) | (0.035) | (0.032) | (0.045) | (0.042) | (0.045) |
| $R_{ik,t-1}^{births}$ | 0.379 | 0.268 | 0.344 | 0.258 | 0.271 | 0.325 |
| | (0.302) | (0.345) | (0.393) | (0.384) | (0.324) | (0.438) |
| FE: Broad categ.-region-period | Y | Y | Y | Y | Y | Y |
| FE: Category-period | Y | Y | Y | Y | Y | Y |
| $\rho_{ik,t-1}^{M}$ * period | Y | Y | Y | Y | Y | Y |
| $\rho_{ik,t-1}^{\omega}$ * period | Y | Y | Y | Y | Y | Y |
| Observations | 3944 | 3944 | 3944 | 3944 | 3944 | 3944 |
| Pseudo-$R^2$ | 0.216 | 0.218 | 0.219 | 0.218 | 0.218 | 0.220 |
| BIC | 9585.9 | 9595.9 | 9597.1 | 9603.6 | 9603.6 | 9609.8 |

Standard errors are clustered by region and period. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S20***. *Logistic regression model explaining exits from existing areas of specializations, interacting measures of spatial proximity with period fixed-effects.*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immigrants}$ | | -0.611*** | -0.592*** | -0.611*** | -0.594*** | -0.596*** |
| | | (0.094) | (0.113) | (0.133) | (0.115) | (0.171) |
| $M_{ik,t-1}^{emigrants}$ | | 0.241 | 0.246 | 0.240 | 0.249 | 0.319 |
| | | (0.325) | (0.326) | (0.328) | (0.470) | (0.390) |
| $\omega_{ik,t-1}^{immigrants}$ | | | -0.084*** | | | -0.085*** |
| | | | (0.013) | | | (0.028) |
| $\omega_{ik,t-1}^{emigrants}$ | | | | -0.0008 | | 0.055 |
| | | | | (0.018) | | (0.041) |
| $\omega_{ik,t-1}^{births}$ | | | | | -0.041** | -0.061 |
| | | | | | (0.020) | (0.049) |
| $ubiquity_{k,t-1}$ | -0.076*** | -0.079*** | -0.078*** | -0.079*** | -0.078*** | -0.081*** |
| | (0.014) | (0.016) | (0.016) | (0.015) | (0.015) | (0.015) |
| $\rho_{ik,t-1}^{M}$ | 6.551** | 7.007** | 6.991** | 7.001** | 7.025*** | 7.446*** |
| | (2.480) | (2.685) | (2.863) | (2.785) | (2.667) | (2.705) |
| $\rho_{ik,t-1}^{\omega}$ | 0.394*** | 0.433*** | 0.463*** | 0.432*** | 0.434*** | 0.482*** |
| | (0.074) | (0.078) | (0.091) | (0.076) | (0.084) | (0.098) |
| $R_{ik,t-1}^{births}$ | 0.032 | 0.032 | 0.015 | 0.032 | 0.027 | 0.018 |
| | (0.026) | (0.041) | (0.023) | (0.024) | (0.027) | (0.021) |
| FE: Broad categ.-region-period | Y | Y | Y | Y | Y | Y |
| FE: Category-period | Y | Y | Y | Y | Y | Y |
| $\rho_{ik,t-1}^{M}$ * period | Y | Y | Y | Y | Y | Y |
| $\rho_{ik,t-1}^{\omega}$ * period | Y | Y | Y | Y | Y | Y |
| Observations | 1051 | 1051 | 1051 | 1051 | 1051 | 1051 |
| Pseudo-$R^2$ | 0.240 | 0.248 | 0.256 | 0.248 | 0.249 | 0.258 |
| BIC | 3644.2 | 3647.6 | 3642.1 | 3654.5 | 3653.2 | 3654.0 |

Standard errors are clustered by region and period. * p < 0.1, ** p < 0.05, *** p < 0.01

### 3.4.4. Excluding the 20th century, and exploring heterogeneous effects across time

We established in Table S3 that our dataset is unbalanced with respect to the different centuries. The majority of famous individuals in our dataset are born in the 19th and 20th century. Hence, observations of relatedness densities and entries or exits are also unbalanced. To check the robustness of our results, we run the logistic regression models excluding the 20th century. Table S21 shows the results explaining entries to new activities, Table S22 for exits of activities. For both entries and exits we find that our main results remain unchanged. That is, $M_{ik,t-1}^{immi}$ and $\omega_{ik,t-1}^{immi}$ correlate positively with future entries and negatively with future exits.

Also, we provide the results for including only the 20th century in Table S23 (entries) and Table S24 (exits). Comparing the coefficients for the 11th to 19th century and the 20th century, we find that the related knowledge of immigrants plays a more significant role in the 20th century than before. Spillovers of migrants within the same activity, however, play a smaller role in the 20th century. Exploring differences across time periods further may be an interesting avenue for future research, since the cost of migration changed substantially over the past centuries.

***Table S21***. *Logistic regression model explaining entries to new activities, subsample for 11th to 19th century.*

| | Dependent Variable: $Entry_{ik,t}$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| $M_{ik,t-1}^{immi}$ | 0.487** | 0.461* | 0.461** | 0.461** | 0.447* |
| | (0.168) | (0.197) | (0.175) | (0.181) | (0.230) |
| $M_{ik,t-1}^{emi}$ | -0.064 | -0.092 | -0.160 | -0.188 | -0.389 |
| | (0.395) | (0.398) | (0.461) | (0.505) | (0.886) |
| $\omega_{ik,t-1}^{immi}$ | 0.009** | 0.015* | 0.014*** | 0.018** | 0.026* |
| | (0.002) | (0.007) | (0.005) | (0.008) | (0.015) |
| $\omega_{ik,t-1}^{emi}$ | 0.0006 | -0.012 | -0.010 | -0.019** | -0.023 |
| | (0.004) | (0.008) | (0.007) | (0.009) | (0.022) |
| $\omega_{ik,t-1}^{births}$ | 0.008 | 0.016 | 0.019** | 0.028*** | 0.022 |
| | (0.009) | (0.009) | (0.008) | (0.010) | (0.019) |
| $diversity_{i,t-1}$ | -0.034 | -0.146** | -0.164*** | | |
| | (0.034) | (0.039) | (0.050) | | |
| $ubiquity_{k,t-1}$ | -0.006 | 0.001 | 0.003 | 0.002 | 0.004 |
| | (0.003) | (0.004) | (0.011) | (0.012) | (0.019) |
| $\rho_{ik,t-1}^{M}$ | 1.125* | 0.901* | 0.564 | 0.637 | 0.083 |
| | (0.437) | (0.373) | (0.737) | (0.761) | (1.212) |
| $\rho_{ik,t-1}^{\omega}$ | -0.010 | -0.032** | -0.017 | -0.018 | 0.036 |
| | (0.019) | (0.007) | (0.034) | (0.039) | (0.068) |
| $R_{ik,t-1}^{births}$ | 0.585 | 0.546 | 0.773** | 0.802** | 0.719 |
| | (0.448) | (0.550) | (0.342) | (0.376) | (0.744) |
| log $(pop_{i,t})$ | -0.024 | 0.233 | 0.238 | | |
| | (0.012) | (0.259) | (0.230) | | |
| FE: period | Y | Y | Y | | |
| FE: region | | Y | Y | | |
| FE: category | | | Y | Y | |
| FE: period-region | | | | Y | |
| FE: region-period-broad category | | | | | Y |
| FE: category-period | | | | | Y |
| Observations | 1397 | 1386 | 1386 | 1382 | 883 |
| Pseudo-$R^2$ | 0.028 | 0.058 | 0.078 | 0.089 | 0.174 |
| BIC | 1803.6 | 2126.3 | 2215.5 | 2409.6 | 2601.6 |

Standard errors are clustered by region. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S22***. *Logistic regression model explaining exits of activities, subsample for 11th to 19th century.*

| | Dependent Variable: $Exit_{ik,t}$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| $M_{ik,t-1}^{immi}$ | -0.431** | -0.448** | -0.445** | -0.548** | -0.635 |
| | (0.169) | (0.191) | (0.213) | (0.223) | (0.689) |
| $M_{ik,t-1}^{emi}$ | -0.233 | -0.220 | -0.115 | -0.225 | -0.284 |
| | (0.253) | (0.287) | (0.276) | (0.299) | (0.791) |
| $\omega_{ik,t-1}^{immi}$ | -0.431** | -0.448** | -0.445** | -0.548** | -0.635 |
| | (0.008) | (0.012) | (0.012) | (0.017) | (0.038) |
| $\omega_{ik,t-1}^{emi}$ | 0.010 | 0.015 | 0.015 | -0.005 | -0.076 |
| | (0.008) | (0.015) | (0.015) | (0.022) | (0.082) |
| $\omega_{ik,t-1}^{births}$ | -0.016 | -0.028 | -0.027 | -0.031 | 0.018 |
| | (0.013) | (0.018) | (0.020) | (0.027) | (0.077) |
| $diversity_{i,t-1}$ | 0.070 | 0.336** | 0.310** | | |
| | (0.066) | (0.138) | (0.138) | | |
| $ubiquity_{k,t-1}$ | 0.007 | 0.011 | 0.010 | 0.024 | 0.047 |
| | (0.018) | (0.021) | (0.026) | (0.030) | (0.064) |
| $\rho_{ik,t-1}^{M}$ | 0.659 | 0.423 | 0.464 | 0.151 | 1.074 |
| | (1.387) | (1.564) | (1.717) | (1.831) | (3.816) |
| $\rho_{ik,t-1}^{\omega}$ | -0.025 | -0.050 | -0.047 | -0.056 | -0.133 |
| | (0.029) | (0.040) | (0.047) | (0.066) | (0.141) |
| $R_{ik,t-1}^{births}$ | 0.001 | -0.005 | -0.012 | -0.002 | -0.023 |
| | (0.037) | (0.057) | (0.063) | (0.067) | (0.119) |
| log $(pop_{i,t})$ | -0.141 | -1.478** | -1.513*** | | |
| | (0.107) | (0.572) | (0.531) | | |
| FE: period | Y | Y | Y | | |
| FE: region | | Y | Y | | |
| FE: category | | | Y | Y | |
| FE: period-region | | | | Y | |
| FE: region-period-broad category | | | | | Y |
| FE: category-period | | | | | Y |
| Observations | 556 | 556 | 550 | 522 | 253 |
| Pseudo-$R^2$ | 0.033 | 0.094 | 0.111 | 0.137 | 0.266 |
| BIC | 846.3 | 1140.5 | 1213.8 | 1281.5 | 960.2 |

Standard errors are clustered by region. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S23.*** *Logistic regression model explaining entries to activities, subsample for 20th century.*

| | Dependent Variable: $Entry_{ik,t}$ | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| $M_{ik,t-1}^{immi}$ | 0.232** | 0.237** | 0.288** | 0.257 |
| | (0.103) | (0.111) | (0.113) | (0.158) |
| $M_{ik,t-1}^{emi}$ | 0.248 | 0.188 | -0.042 | 0.139 |
| | (0.267) | (0.312) | (0.326) | (0.423) |
| $\omega_{ik,t-1}^{immi}$ | 0.016*** | 0.020* | 0.027** | 0.037** |
| | (0.004) | (0.011) | (0.012) | (0.017) |
| $\omega_{ik,t-1}^{emi}$ | 0.010 | -0.046 | -0.044 | -0.023 |
| | (0.013) | (0.030) | (0.032) | (0.049) |
| $\omega_{ik,t-1}^{births}$ | 0.012 | 0.046 | 0.045 | 0.025 |
| | (0.017) | (0.029) | (0.032) | (0.046) |
| $diversity_{i,t-1}$ | -0.026 | | | |
| | (0.026) | | | |
| $ubiquity_{k,t-1}$ | -0.006 | -0.013** | -0.003 | 0.001 |
| | (0.005) | (0.006) | (0.006) | (0.008) |
| $\rho_{ik,t-1}^{M}$ | 1.375 | 2.380** | 1.491 | 0.863 |
| | (1.050) | (1.178) | (1.221) | (1.714) |
| $\rho_{ik,t-1}^{\omega}$ | 0.113*** | 0.180*** | 0.113** | 0.122* |
| | (0.033) | (0.043) | (0.049) | (0.065) |
| $R_{ik,t-1}^{births}$ | 0.300 | 0.314 | 0.498* | 0.246 |
| | (0.226) | (0.260) | (0.276) | (0.372) |
| log $(pop_{i,t})$ | 0.206*** | | | |
| | (0.042) | | | |
| FE: region | | Y | Y | |
| FE: category | | | Y | Y |
| FE: region-broad category | | | | Y |
| Observations | 4783 | 4783 | 4783 | 3061 |
| Pseudo-$R^2$ | 0.072 | 0.100 | 0.148 | 0.213 |
| BIC | 4011.4 | 4953.5 | 4946.0 | 6534.4 |

Standard errors are clustered by region. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S24.*** *Logistic regression model explaining exits of activities, subsample for 20th century.*

| | Dependent Variable: $Exit_{ik,t}$ | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| $M_{ik,t-1}^{immi}$ | -0.487*** | -0.511*** | -0.540*** | -0.625** |
| | (0.145) | (0.160) | (0.172) | (0.297) |
| $M_{ik,t-1}^{emi}$ | 0.277 | 0.270 | 0.347 | 0.465 |
| | (0.202) | (0.228) | (0.249) | (0.418) |
| $\omega_{ik,t-1}^{immi}$ | -0.012** | -0.051*** | -0.058** | -0.098** |
| | (0.006) | (0.018) | (0.023) | (0.046) |
| $\omega_{ik,t-1}^{emi}$ | 0.011 | 0.034 | 0.020 | 0.079 |
| | (0.016) | (0.031) | (0.033) | (0.063) |
| $\omega_{ik,t-1}^{births}$ | 0.019 | -0.038 | -0.027 | -0.104 |
| | (0.018) | (0.037) | (0.039) | (0.075) |
| $diversity_{i,t-1}$ | -0.082** | | | |
| | (0.041) | | | |
| $ubiquity_{k,t-1}$ | -0.014* | -0.030*** | -0.043*** | -0.091*** |
| | (0.007) | (0.008) | (0.010) | (0.023) |
| $\rho_{ik,t-1}^{M}$ | 0.390 | 1.931 | 1.855 | 9.130* |
| | (1.510) | (1.931) | (2.169) | (4.940) |
| $\rho_{ik,t-1}^{\omega}$ | 0.021 | 0.184*** | 0.256*** | 0.555*** |
| | (0.048) | (0.060) | (0.079) | (0.165) |
| $R_{ik,t-1}^{births}$ | -0.002 | 0.000 | -0.012 | 0.033 |
| | (0.016) | (0.021) | (0.024) | (0.044) |
| log $(pop_{i,t})$ | -0.240** | | | |
| | (0.093) | | | |
| FE: region | | Y | Y | |
| FE: category | | | Y | Y |
| FE: region-broad category | | | | Y |
| Observations | 1489 | 1467 | 1461 | 798 |
| Pseudo-$R^2$ | 0.065 | 0.131 | 0.181 | 0.233 |
| BIC | 1824.5 | 2575.8 | 2640.6 | 2457.8 |

Standard errors are clustered by region. * p < 0.1, ** p < 0.05, *** p < 0.01

### 3.4.5. Redefining entries and exits

In the main results, we define entries based on births in a location in the coming century. That is, $Entry_{ik,t} = 1$ if $M_{ik,t-1}^{births} = 0$ and $M_{ik,t}^{births} = 1$ (see Eq. S2). Similarly, we defined $Exit_{ik,t} = 1$ if $M_{ik,t-1}^{births} = 1$ and $M_{ik,t}^{births} = 0$

To check the robustness of our results, we apply a different definition of entries and exits. Instead of defining the entry to a new activity by developing a new specialization, we can define entry as a location exhibiting births of famous individuals with a certain occupation for the first time. Specifically, let us refer to this definition of entry as $Entry2_{ik,t}$ and let it be defined as $Entry2_{ik,t} = 1$ if $N_{ik,t-1}^{births} = 0$ and $N_{ik,t}^{births} > 0$. Similarly, we can define exits as $Exit2_{ik,t} = 1$ if $N_{ik,t-1}^{births} > 0$ and $N_{ik,t}^{births} = 0$.

Table S25 shows the results of the logistic regression model for this definition of entry, for different specifications of fixed effects. As can be seen, the results are robust, that is, $M_{ik,t-1}^{immi}$ and $\omega_{ik,t-1}^{immi}$ correlate positively with future entries. The same holds for redefining exits as last births of famous individuals with a certain activity (Table S26). While $M_{ik,t-1}^{immi}$ correlates significantly for all fixed effects, the diffusion of related knowledge is – in contrast to the main findings defining entries and exits as gaining or losing a specialization – not significant for the most restrictive specifications.

***Table S25**. Regression results explaining entries to new activities, redefining entries as first births of famous individuals with occupation k in location i*

| | Dependent Variable: $Entry2_{ik,t}$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| $M_{ik,t-1}^{immi}$ | 0.268** | 0.271** | 0.308*** | 0.346*** | 0.376** |
| | (0.125) | (0.125) | (0.099) | (0.118) | (0.154) |
| $\omega_{ik,t-1}^{immi}$ | 0.021*** | 0.021** | 0.024*** | 0.028*** | 0.011 |
| | (0.004) | (0.006) | (0.004) | (0.005) | (0.009) |
| $\omega_{ik,t-1}^{emi}$ | 0.015*** | 0.006 | 0.004 | -0.012 | 0.011 |
| | (0.002) | (0.007) | (0.007) | (0.023) | (0.029) |
| $\omega_{ik,t-1}^{births}$ | -0.009 | 0.002 | 0.0002 | 0.008 | -0.017 |
| | (0.008) | (0.006) | (0.006) | (0.010) | (0.015) |
| $diversity_{i,t-1}$ | 0.005 | -0.077** | -0.059 | | |
| | (0.023) | (0.035) | (0.036) | | |
| $ubiquity_{k,t-1}$ | 0.003 | 0.004 | 0.00001 | -0.001 | -0.010 |
| | (0.004) | (0.006) | (0.006) | (0.007) | (0.009) |
| $\rho_{ik,t-1}^{M}$ | 2.175*** | 2.320*** | 2.830** | 3.065*** | 3.949** |
| | (0.410) | (0.395) | (1.109) | (1.165) | (1.751) |
| $\rho_{ik,t-1}^{\omega}$ | 0.047 | 0.041 | 0.075** | 0.090** | 0.228** |
| | (0.043) | (0.067) | (0.036) | (0.040) | (0.065) |
| log $(pop_{i,t})$ | 0.279*** | 0.197 | 0.237 | | |
| | (0.069) | (0.166) | (0.199) | | |
| FE: period | Y | Y | Y | | |
| FE: region | | Y | Y | | |
| FE: category | | | Y | Y | |
| FE: period-region | | | | Y | |
| FE: region-period-broad category | | | | | Y |
| FE: category-period | | | | | Y |
| Observations | 5569 | 5539 | 5539 | 5487 | 3536 |
| Pseudo-R2 | 0.168 | 0.225 | 0.317 | 0.320 | 0.384 |
| BIC | 3181.3 | 3998.6 | 3805.8 | 4215.2 | 4630.0 |

* p < 0.1, ** p < 0.05, *** p < 0.01

***Table S26.** Regression results explaining exits to new activities, redefining exits as last births of famous individuals with occupation k in location i*

| | Dependent Variable: $Exit2_{ik,t}$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| $M_{ik,t-1}^{immi}$ | -0.360*** | -0.380*** | -0.416*** | -0.459*** | -0.336** |
| | (0.061) | (0.066) | (0.086) | (0.093) | (0.141) |
| $M_{ik,t-1}^{emi}$ | 0.216 | 0.147 | 0.074 | 0.083 | 0.367 |
| | (0.133) | (0.119) | (0.180) | (0.202) | (0.305) |
| $\omega_{ik,t-1}^{immi}$ | -0.017** | -0.026*** | -0.025*** | -0.025 | -0.058 |
| | (0.007) | (0.010) | (0.006) | (0.017) | (0.037) |
| $\omega_{ik,t-1}^{emi}$ | 0.010 | 0.015* | 0.012* | 0.023 | 0.029 |
| | (0.009) | (0.008) | (0.007) | (0.016) | (0.024) |
| $\omega_{ik,t-1}^{births}$ | 0.007 | -0.016** | -0.018*** | -0.037*** | 0.010 |
| | (0.010) | (0.007) | (0.007) | (0.012) | (0.046) |
| $diversity_{i,t-1}$ | -0.067 | 0.071* | 0.082* | | |
| | (0.038) | (0.031) | (0.033) | | |
| $ubiquity_{k,t-1}$ | -0.015*** | -0.019** | -0.030** | -0.029** | -0.042** |
| | (0.003) | (0.006) | (0.008) | (0.010) | (0.013) |
| $\rho_{ik,t-1}^{M}$ | -1.866** | -2.428** | -1.373** | -1.919** | -3.378 |
| | (0.586) | (1.019) | (0.590) | (0.601) | (2.349) |
| $\rho_{ik,t-1}^{\omega}$ | 0.041 | 0.113* | 0.096 | 0.098 | 0.060 |
| | (0.042) | (0.059) | (0.072) | (0.094) | (0.169) |
| log $(pop_{i,t})$ | -0.358*** | -0.211 | -0.275 | | |
| | (0.036) | (0.345) | (0.360) | | |
| FE: period | Y | Y | Y | | |
| FE: region | | Y | Y | | |
| FE: category | | | Y | Y | |
| FE: period-region | | | | Y | |
| FE: region-period-broad category | | | | | Y |
| FE: category-period | | | | | Y |
| Observations | 2656 | 2656 | 2628 | 2521 | 1488 |
| Pseudo-$R^2$ | 0.168 | 0.225 | 0.317 | 0.320 | 0.384 |
| BIC | 3181.3 | 3998.6 | 3805.8 | 4215.2 | 4630.0 |

* p < 0.1, ** p < 0.05, *** p < 0.01

### 3.4.6. Interaction terms

Following the literature on the role of migration in unrelated diversification (Elekes et al., 2019; Miguelez & Morrison, 2022; Neffke et al., 2018), we add interaction terms between various relatedness densities to the main specification of column 6 of Table S6. For example, a significantly negative interaction term between $\omega_{ik}^{immi}$ and $\omega_{ik}^{births}$ would indicate that the related knowledge of immigrants and those of individuals born in a location are substitutes to each other. Put differently, if $\omega_{ik}^{births}$ is high, the correlation of $\omega_{ik}^{immi}$ with the probability of entry decreases. As Table S27 shows, the interaction term between $\omega_{ik}^{immi}$ and $\omega_{ik}^{births}$ is indeed significantly negative across all fixed-effects specifications. However, quantitatively, the coefficient is very small compared to the overall coefficient $\omega_{ik}^{immi}$. Thus, we cannot conclude that migration contributes substantially to unrelated diversification.

***Table S27****. Regression results explaining entries to new activities, including interaction terms*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| $M_{ik,t-1}^{immi}$ | 0.320*** | 0.319*** | 0.319*** | 0.317*** | 0.316*** | 0.316*** | 0.330*** | 0.329*** | 0.330*** |
| | (0.077) | (0.077) | (0.076) | (0.074) | (0.074) | (0.075) | (0.064) | (0.063) | (0.066) |
| $M_{ik,t-1}^{emi}$ | 0.155 | 0.137 | 0.118 | 0.186 | 0.168 | 0.152 | -0.012 | -0.024 | -0.035 |
| | (0.163) | (0.164) | (0.164) | (0.210) | (0.204) | (0.212) | (0.222) | (0.224) | (0.228) |
| $\omega_{ik,t-1}^{immi}$ | 0.021*** | 0.030*** | 0.011*** | 0.025*** | 0.045*** | 0.010*** | 0.024*** | 0.045*** | 0.011*** |
| | (0.003) | (0.005) | (0.003) | (0.008) | (0.010) | (0.003) | (0.007) | (0.010) | (0.003) |
| $\omega_{ik,t-1}^{emi}$ | 0.013*** | 0.001 | 0.015 | 0.009 | -0.012*** | 0.007 | 0.007* | -0.012*** | 0.003 |
| | (0.005) | (0.004) | (0.009) | (0.007) | (0.004) | (0.014) | (0.003) | (0.005) | (0.009) |
| $\omega_{ik,t-1}^{births}$ | 0.003 | 0.024*** | 0.019 | 0.011 | 0.050*** | 0.029 | 0.009 | 0.047*** | 0.023* |
| | (0.005) | (0.004) | (0.012) | (0.007) | (0.013) | (0.020) | (0.006) | (0.010) | (0.012) |
| $\omega_{ik,t-1}^{immi} * \omega_{ik,t-1}^{emi}$ | -0.000*** | | | -0.000* | | | -0.000** | | |
| | (0.000) | | | (0.000) | | | (0.000) | | |
| $\omega_{ik,t-1}^{immi} * \omega_{ik,t-1}^{births}$ | | -0.001*** | | | -0.001*** | | | -0.001*** | |
| | | (0.000) | | | (0.000) | | | (0.000) | |
| $\omega_{ik,t-1}^{emi} * \omega_{ik,t-1}^{births}$ | | | 0.000 | | | 0.000 | | | 0.000 |
| | | | (0.000) | | | (0.000) | | | (0.000) |
| $diversity_{i,t-1}$ | 0.008 | 0.004 | 0.003 | -0.030 | -0.025 | -0.032 | -0.022 | -0.017 | -0.024 |
| | (0.013) | (0.012) | (0.018) | (0.021) | (0.024) | (0.022) | (0.022) | (0.025) | (0.022) |
| $ubiquity_{k,t-1}$ | 0.006** | 0.006** | 0.007** | 0.008** | 0.007** | 0.008** | 0.006*** | 0.006*** | 0.007*** |
| | (0.003) | (0.003) | (0.003) | (0.004) | (0.004) | (0.004) | (0.001) | (0.001) | (0.001) |
| $\rho_{ik,t-1}^{M}$ | 0.232 | 0.163 | 0.096 | 0.302 | 0.226*** | 0.163*** | 0.621*** | 0.545*** | 0.514*** |
| | (0.340) | (0.292) | (0.220) | (0.193) | (0.086) | (0.054) | (0.146) | (0.053) | (0.146) |
| $\rho_{ik,t-1}^{\omega}$ | 0.026 | 0.027 | 0.028 | 0.013 | 0.012 | 0.016 | 0.016 | 0.015 | 0.018 |
| | (0.033) | (0.033) | (0.032) | (0.041) | (0.040) | (0.041) | (0.024) | (0.023) | (0.024) |
| $R_{ik,t-1}^{births}$ | 0.255* | 0.288* | 0.275* | 0.222 | 0.259 | 0.234 | 0.531*** | 0.557*** | 0.530*** |
| | (0.152) | (0.155) | (0.147) | (0.169) | (0.169) | (0.170) | (0.185) | (0.188) | (0.184) |
| log $(pop_{i,t})$ | 0.138* | 0.144** | 0.143* | 0.249*** | 0.222*** | 0.265*** | 0.242*** | 0.218** | 0.255*** |
| | (0.072) | (0.070) | (0.078) | (0.068) | (0.081) | (0.079) | (0.081) | (0.089) | (0.087) |
| FE: period | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| FE: region | | | | Y | Y | Y | Y | Y | Y |
| FE: occu. category | | | | | | | Y | Y | Y |
| Observations | 6180 | 6180 | 6180 | 6180 | 6180 | 6180 | 6180 | 6180 | 6180 |
| Pseudo-$R^2$ | 0.076 | 0.077 | 0.076 | 0.097 | 0.099 | 0.097 | 0.125 | 0.127 | 0.125 |
| AIC | 5671.9 | 5666.3 | 5670.7 | 5802.4 | 5789.8 | 5802.3 | 5678.6 | 5666.8 | 5679.5 |
| BIC | 5793.0 | 5787.5 | 5791.8 | 6798.3 | 6785.7 | 6798.2 | 6836.0 | 6824.2 | 6836.9 |

Standard errors are clustered by period and region. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

### 3.4.7. Heterogenous effects across activities

The effects of knowledge spillovers may differ across activities. The emergence of artists or scientists may be more demand-driven than other categories, since it is well known that wealthy patrons supported the arts and sciences. The same may apply to sports persons. The emergence of politicians may be less related to the presence of artists or sports persons but is affected by political decisions. Businessmen may profit from spillovers across the sciences and institutions.

To explore these heterogenous effects, we split our dataset into six different, highly aggregated occupational categories:

(1) "Arts",

(2) "Humanities",

(3) "Sciences",

(4) "Business & Technology" (this category includes the category 'Business & Law' as well as engineers and inventors),

(5) "Sports", and

(6) "Public Figures & Institutions".

This distinction loosely follows the full taxonomy (Table S2).

For each of these categories, we run logistic regression models for entries and exits to estimate whether the role of immigrants, emigrants and locals differs across these categories.

The results are shown in Table S28 and Table S29.

For entries (Table S28), we find that the knowledge of immigrants in the same activity ($M_{ik,t-1}^{immi}$) correlates with future entries in "Sciences" as well as for "Public Figures & Institutions". That is, the probability of entering e.g. physics or politics increases with a disproportionate inflow of physicists or politicians. The related knowledge of immigrants ($\omega_{ik,t-1}^{immi}$) correlates positively with entries in "Sports" and "Humanities". That is, the emergence of sportsmen, writers or philosophers correlates with the immigration of individuals in related activities.

For exits (Table S29), we find that the probability of exit decreases with the knowledge of immigrants in the same activity ($M_{ik,t-1}^{immi}$) in "Arts" and "Public Figures & Institutions". A larger than expected inflow of e.g. painters decreases the probability of exiting painting. The related knowledge of immigrants ($\omega_{ik,t-1}^{immi}$) correlates negatively with exits in "Business &

Technology". The immigration of individuals with related activities decreases the probability of exit occupations in business or technology.

These models help understand the mechanisms behind our findings. The number of observations, however, is more limited in these models than in our full model and, hence, we cannot be as restrictive with fixed effects. More comprehensive data would be required to analyze the heterogeneity across different fields in more detail and to provide more robust evidence on the specific mechanisms. Exploring the heterogeneity across disciplines further may be an interesting avenue for future research.

***Table S28.*** *Heterogeneous effects for entries across occupation categories*

| | Dependent Variable: $Entry_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | Arts | Humanities | Sciences | Business & Technology | Sports | Public Figures & Institutions |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | 0.077 | 0.304 | 0.438** | -0.665 | -0.158 | 0.576*** |
| | (0.151) | (0.393) | (0.125) | (0.545) | (0.563) | (0.094) |
| $M_{ik,t-1}^{emi}$ | 0.035 | -0.343 | -0.851 | -1.227 | | 0.625 |
| | (0.714) | (0.543) | (0.491) | (1.652) | | (0.770) |
| $\omega_{ik,t-1}^{immi}$ | 0.008 | 0.031** | 0.015 | -0.020 | 0.163*** | 0.020 |
| | (0.019) | (0.012) | (0.012) | (0.019) | (0.060) | (0.014) |
| $\omega_{ik,t-1}^{emi}$ | -0.004 | -0.011 | 0.004 | -0.039 | 0.043 | -0.009 |
| | (0.031) | (0.020) | (0.022) | (0.029) | (0.094) | (0.022) |
| $\omega_{ik,t-1}^{births}$ | 0.038 | 0.005 | 0.007 | -0.015 | -0.061 | 0.016 |
| | (0.025) | (0.025) | (0.033) | (0.021) | (0.103) | (0.032) |
| $diversity_{i,t-1}$ | -0.058 | -0.051 | -0.011 | 0.188 | | -0.128 |
| | (0.050) | (0.079) | (0.065) | (0.103) | | (0.084) |
| $ubiquity_{k,t-1}$ | -0.001 | 0.009 | 0.024** | 0.173*** | 0.041 | 0.007 |
| | (0.009) | (0.009) | (0.006) | (0.027) | (0.036) | (0.008) |
| $\rho_{ik,t-1}^{M}$ | -0.654 | 3.744** | -1.147 | 2.131 | -2.384 | 1.170 |
| | (1.592) | (1.641) | (1.067) | (2.097) | (2.142) | (1.717) |
| $\rho_{ik,t-1}^{\omega}$ | 0.131 | -0.198*** | 0.068 | -0.446** | -0.082 | 0.034 |
| | (0.082) | (0.068) | (0.044) | (0.102) | (0.238) | (0.033) |
| $R_{ik,t-1}^{births}$ | 0.188 | 0.134 | 0.383 | 1.625 | 2.569 | 1.251*** |
| | (0.297) | (0.842) | (0.193) | (2.136) | (2.993) | (0.286) |
| log $(pop_{i,t})$ | 0.277 | 0.340 | 0.808** | 0.196 | | 0.700 |
| | (0.191) | (0.227) | (0.245) | (0.948) | | (0.421) |
| FE: period | Y | Y | Y | Y | Y | Y |
| FE: region | Y | Y | Y | Y | Y | Y |
| FE: category | Y | Y | Y | Y | Y | Y |
| Num.Obs. | 1502 | 515 | 1094 | 288 | 521 | 939 |
| Pseudo-R2 | 0.194 | 0.199 | 0.179 | 0.248 | 0.210 | 0.246 |
| BIC | 2316.1 | 1146.5 | 1807.0 | 665.1 | 1044.0 | 1581.4 |

Standard errors are clustered by period and region. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

***Table S29**. Heterogeneous effects for exits across occupation categories*

| | Dependent Variable: $Exit_{ik,t}$ | | | | | |
|---|---|---|---|---|---|---|
| | Arts | Humanities | Sciences | Business & Technology | Sports | Public Figures & Institutions |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $M_{ik,t-1}^{immi}$ | -1.348** | -0.929 | 0.031 | -3.827 | -1.335 | -0.218* |
| | (0.426) | (0.762) | (0.211) | (3.747) | (1.139) | (0.083) |
| $M_{ik,t-1}^{emi}$ | -0.214 | -1.115 | 0.211 | 4.889* | | -0.533 |
| | (0.243) | (0.816) | (0.243) | (2.349) | | (0.615) |
| $\omega_{ik,t-1}^{immi}$ | -0.026 | 0.071 | -0.007 | -0.811** | -0.610 | -0.006 |
| | (0.023) | (0.046) | (0.020) | (0.363) | (0.455) | (0.011) |
| $\omega_{ik,t-1}^{emi}$ | -0.031 | -0.023 | 0.019 | -0.622 | | 0.002 |
| | (0.015) | (0.036) | (0.049) | (0.479) | | (0.031) |
| $\omega_{ik,t-1}^{births}$ | -0.017 | 0.031 | -0.028 | 0.486 | 0.406 | 0.042 |
| | (0.047) | (0.037) | (0.057) | (0.555) | (0.291) | (0.023) |
| $diversity_{i,t-1}$ | 0.231* | -0.458* | -0.084 | -2.513** | | -0.233 |
| | (0.085) | (0.259) | (0.121) | (1.060) | | (0.133) |
| $ubiquity_{k,t-1}$ | -0.030 | -0.052*** | -0.027 | -0.102 | -0.886** | -0.014 |
| | (0.036) | (0.018) | (0.017) | (0.300) | (0.341) | (0.011) |
| $\rho_{ik,t-1}^{M}$ | 2.116 | 4.253 | 3.286 | 68.536 | 70.545 | 0.622 |
| | (7.694) | (4.700) | (1.690) | (45.770) | (48.115) | (2.932) |
| $\rho_{ik,t-1}^{\omega}$ | 0.203 | 0.083 | -0.052 | 0.508 | 4.863* | -0.285*** |
| | (0.169) | (0.187) | (0.151) | (2.412) | (2.386) | (0.043) |
| $R_{ik,t-1}^{births}$ | 0.017 | -0.089 | 0.005 | 0.118 | 0.490** | -0.212* |
| | (0.045) | (0.088) | (0.030) | (0.195) | (0.204) | (0.084) |
| log $(pop_{i,t})$ | -0.505 | 2.274 | -0.891 | 2.534 | | -0.772 |
| | (0.646) | (1.643) | (0.766) | (6.673) | | (0.398) |
| FE: period | Y | Y | Y | Y | Y | Y |
| FE: region | Y | Y | Y | Y | Y | Y |
| FE: category | Y | Y | Y | | Y | Y |
| Num.Obs. | 360 | 159 | 435 | 63 | 46 | 323 |
| Pseudo-R2 | 0.239 | 0.253 | 0.201 | 0.662 | 0.592 | 0.209 |
| BIC | 884.8 | 494.0 | 1003.9 | 170.3 | 117.7 | 862.0 |

Standard errors are clustered by period and region. We cannot include category fixed-effects in column (4), since the maximum likelihood estimator does not converge if they are included. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

### 3.4.8. Heterogenous effects across city size

It may be that the patterns we discover here vary for cities of different sizes. Hence, we explore this potential heterogeneity by splitting the sample into small cities (population levels by Buringh (2021) below the median of the respective century) and large cities (population levels above the median of the respective century).

Table S30 shows the results for entries and exits for both, small and large cities. Indeed, the correlational patterns we uncover in this study vary across sizes. The knowledge of immigrants in the same activity and in related activities is significantly more relevant for large cities with respect to both, entries and exits. In contrast, it can be seen that the knowledge of emigrants is a highly relevant factor in predicting future exits of activities for small cities.

***Table S30**. Heterogeneous effects across city size for entries and exits*

| | Dependent Variable: $Entry_{ik,t}$ | | Dependent Variable: $Exit_{ik,t}$ | |
|---|---|---|---|---|
| | Small cities | Large cities | Small cities | Large cities |
| | (1) | (2) | (3) | (4) |
| $M_{ik,t-1}^{immi}$ | 0.042 | 0.310*** | -1.065* | -0.576*** |
| | (0.242) | (0.085) | (0.607) | (0.172) |
| $M_{ik,t-1}^{emi}$ | -0.581 | -0.019 | 2.519*** | 0.005 |
| | (0.323) | (0.350) | (0.219) | (0.228) |
| $\omega_{ik,t-1}^{immi}$ | -0.003 | 0.030*** | -0.122 | -0.052*** |
| | (0.038) | (0.006) | (0.089) | (0.008) |
| $\omega_{ik,t-1}^{emi}$ | 0.066 | -0.027 | 0.214 | -0.043 |
| | (0.057) | (0.021) | (0.275) | (0.064) |
| $\omega_{ik,t-1}^{births}$ | -0.090 | 0.036** | -0.080 | -0.025 |
| | (0.061) | (0.016) | (0.063) | (0.044) |
| $ubiquity_{k,t-1}$ | 0.008 | 0.006** | -0.072** | -0.051*** |
| | (0.008) | (0.003) | (0.023) | (0.013) |
| $\rho_{ik,t-1}^{M}$ | -2.128 | -0.140 | 6.702*** | 5.876*** |
| | (2.091) | (0.663) | (0.115) | (1.294) |
| $\rho_{ik,t-1}^{\omega}$ | 0.208** | 0.073* | 0.087 | 0.153 |
| | (0.049) | (0.044) | (0.344) | (0.174) |
| $R_{ik,t-1}^{births}$ | 2.741*** | 0.096 | -0.071 | 0.018 |
| | (0.550) | (0.241) | (0.054) | (0.035) |
| FE: region-period-broad category | Y | Y | Y | Y |
| FE: category-period | Y | Y | Y | Y |
| Num.Obs. | 906 | 2994 | 142 | 899 |
| Pseudo-$R^2$ | 0.254 | 0.215 | 0.285 | 0.246 |

The number of observations is not equal for small and large cities in spite of splitting the sample at the median of population levels, because smaller cities experience less entries of new and exits of existing activities than large cities do. Hence, many observations for small cities are be removed due to no variety within the fixed-effects structure. Standard errors are clustered by period and region.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

### 3.4.9. Marginal effects after decomposing RCA values

Taking the ratio of the observed and expected number throughout the study, e.g. by considering the Revealed Comparative Advantage / Location Quotient, is first and foremost a control of size. This makes it possible to sensibly compare Paris and London with East Wales and Lower Austria. However, these models are also opaque, not telling us whether our results are driven by changes in the observed or expected number (or both).

In this chapter, we provide results for entries and exits while including all terms of the original ratio. That is, we include the terms $\sum_k N_{ik,t}^{births}$ and $\sum_i N_{ik,t}^{births}$ from the dependent variables $Entry_{ik,t}$ and $Entry_{ik,t}$. We further decompose the terms $M_{ik,t-1}^{immi}$ and $M_{ik,t-1}^{emi}$ into $N_{ik,t-1}^{immi}$, $\sum_k N_{ik,t-1}^{immi}$, $\sum_i N_{ik,t-1}^{immi}$ and $N_{ik,t-1}^{emi}$, $\sum_k N_{ik,t-1}^{emi}$, $\sum_i N_{ik,t-1}^{emi}$, respectively. The coefficients of the observed values, i.e. $N_{ik,t-1}^{immi}$ and $N_{ik,t-1}^{emi}$, can then more directly be interpreted as the marginal effects of one additional immigrant or emigrant with a specific occupation. To reduce skew, all these terms enter transformed using the inverted hyperbolic sine function (denoted by *asinh*). We are using this instead of a log transformation, because we have observations with zeros.

Table S31 and Table S32 show the results for entries and exits, respectively, for various fixed effects specifications. The number of immigrants with a specific occupation ($N_{ik,t-1}^{immi}$) correlates

positively with future entries and negatively with future exits, confirming our main results with composite indices. Using columns (4), we can assess the average marginal effects of one additional immigrant to a region with a certain occupation. We calculate the average marginal effect by comparing the predicted values of the model in column (4) to the predicted values if $N_{ik,t-1}^{immi}$ was increased by 1 for each observation (before using the inverted hyperbolic sine transformation). We find that this average marginal effect amounts to 1.68 percentage points for entries and -5.04 percentage points for exits.

We can also interpret the results as elasticities. We can directly use the coefficients of the model in column (4) to find that a 1 percent increase in $N_{ik,t-1}^{immi}$ translates into an increase of the odds ratio to enter by 0.16 percent ($\exp(0.16 * 0.01)$) and a decrease of the odds ratio to exit by 0.413 percent ($\exp(-0.414 * 0.01)$). To calculate elasticities with respect to the probability of entry or exit, we calculate the average marginal effect by comparing the predicted values of the model in column (4) to the predicted values if $N_{ik,t-1}^{immi}$ was increased by 1% for each observation (before using the inverted hyperbolic sine transformation). We find that a 1% increase in $N_{ik,t-1}^{immi}$ increases the probability of entry by 0.0053% and reduces the probability of exit by 0.033%.

Table S33 and Table S34 show the results for our second definition of entries and exits using the first and last births of individuals with a certain activity in a region (see SM chapter 0). Using again columns (4), we find an average marginal effect of one additional immigrant in a region with a certain activity of 3.38 percentage points for entries and -4.23 percentage points for exits.

Also, in all these regressions the results for the related knowledge of immigrants, emigrants and locals remain virtually unchanged compared to our original results.

***Table S31.*** *Logistic regression models explaining entries, decomposing RCA values*

| | Dependent Variable: $Entry_{ik,t}$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| asinh ($N_{ik,t-1}^{immi}$) | 0.127* | 0.129* | 0.140** | 0.160** | 0.040 |
| | (0.070) | (0.068) | (0.069) | (0.077) | (0.078) |
| asinh ($\sum_k N_{ik,t-1}^{immi}$) | -0.087** | -0.135 | -0.131 | | |
| | (0.041) | (0.141) | (0.137) | | |
| asinh ($\sum_i N_{ik,t-1}^{immi}$) | 0.108 | 0.081 | -0.277 | -0.358 | -0.375 |
| | (0.660) | (0.667) | (0.544) | (0.581) | (0.610) |
| asinh ($N_{ik,t-1}^{emi}$) | -0.151** | -0.166** | 0.001 | -0.042 | 0.050 |
| | (0.062) | (0.053) | (0.028) | (0.051) | (0.047) |
| asinh ($\sum_k N_{ik,t-1}^{emi}$) | 0.010 | 0.002 | -0.011 | | |
| | (0.107) | (0.185) | (0.187) | | |
| asinh ($\sum_i N_{ik,t-1}^{emi}$) | -0.550 | -0.517 | -0.079 | -0.034 | 0.053 |
| | (0.569) | (0.580) | (0.515) | (0.541) | (0.623) |
| asinh ($\sum_k N_{ik,t}^{births}$) | 0.349*** | 0.305** | 0.299* | | |
| | (0.041) | (0.144) | (0.160) | | |
| asinh ($\sum_i N_{ik,t}^{births}$) | 0.546*** | 0.554*** | 0.582*** | 0.591*** | 0.659*** |
| | (0.035) | (0.038) | (0.048) | (0.052) | (0.084) |
| $\omega_{ik,t-1}^{immi}$ | 0.012*** | 0.014*** | 0.014*** | 0.020** | 0.033** |
| | (0.003) | (0.003) | (0.003) | (0.007) | (0.013) |
| $\omega_{ik,t-1}^{emi}$ | 0.007*** | -0.002 | -0.002 | -0.010 | -0.014 |
| | (0.001) | (0.003) | (0.004) | (0.011) | (0.022) |
| $\omega_{ik,t-1}^{births}$ | 0.002 | 0.009 | 0.010 | 0.022** | 0.016 |
| | (0.003) | (0.006) | (0.006) | (0.009) | (0.015) |
| $diversity_{i,t-1}$ | -0.006 | -0.026 | -0.027 | | |
| | (0.022) | (0.022) | (0.021) | | |
| $ubiquity_{k,t-1}$ | 0.011** | 0.012** | 0.008*** | 0.008 | 0.011 |
| | (0.004) | (0.004) | (0.002) | (0.004) | (0.007) |
| $\rho_{ik,t-1}^{M}$ | 0.233 | 0.278 | 0.368*** | 0.476*** | -0.805 |
| | (0.501) | (0.432) | (0.050) | (0.052) | (0.668) |
| $\rho_{ik,t-1}^{\omega}$ | 0.027 | 0.020 | 0.004 | 0.000 | 0.037 |
| | (0.028) | (0.036) | (0.022) | (0.058) | (0.044) |
| log ($pop_{i,t}$) | 0.064 | 0.268* | 0.243* | | |
| | (0.047) | (0.116) | (0.114) | | |
| FE: period | Y | Y | Y | | |
| FE: region | | Y | Y | | |
| FE: category | | | Y | Y | |
| FE: period-region | | | | Y | |
| FE: region-period-broad category | | | | | Y |
| FE: category-period | | | | | Y |
| Observations | 6216 | 6216 | 6216 | 6193 | 3948 |
| Pseudo-R2 | 0.111 | 0.129 | 0.136 | 0.148 | 0.225 |
| BIC | 5670.3 | 6693.6 | 6861.2 | 7487.4 | 9574.5 |

Standard errors are clustered by period and region. *asinh()* denotes the inverted hyperbolic sine function. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S32**. Logistic regression models explaining exits, decomposing RCA values*

| | Dependent Variable: $Exit_{ik,t}$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| asinh ($N_{ik,t-1}^{immi}$) | -0.294*** | -0.323*** | -0.324*** | -0.414*** | -0.378*** |
| | (0.040) | (0.036) | (0.045) | (0.011) | (0.092) |
| asinh ($\sum_k N_{ik,t-1}^{immi}$) | 0.221*** | 0.445** | 0.421** | | |
| | (0.044) | (0.148) | (0.151) | | |
| asinh ($\sum_i N_{ik,t-1}^{immi}$) | 0.150 | 0.407 | 0.533 | 0.510 | -2.546 |
| | (0.516) | (0.537) | (0.702) | (0.967) | (2.330) |
| asinh ($N_{ik,t-1}^{emi}$) | -0.070 | -0.056 | -0.029 | -0.015 | 0.046 |
| | (0.047) | (0.048) | (0.049) | (0.045) | (0.206) |
| asinh ($\sum_k N_{ik,t-1}^{emi}$) | -0.290** | -0.700* | -0.790** | | |
| | (0.111) | (0.315) | (0.296) | | |
| asinh ($\sum_i N_{ik,t-1}^{emi}$) | 0.192 | -0.040 | -0.101 | -0.114 | 3.110 |
| | (0.530) | (0.565) | (0.847) | (1.136) | (2.377) |
| asinh ($\sum_k N_{ik,t}^{births}$) | -0.285** | -0.298 | -0.312 | | |
| | (0.141) | (0.207) | (0.210) | | |
| asinh ($\sum_i N_{ik,t}^{births}$) | -0.499*** | -0.521*** | -0.422*** | -0.479*** | -0.721*** |
| | (0.069) | (0.077) | (0.100) | (0.083) | (0.099) |
| $\omega_{ik,t-1}^{immi}$ | -0.019*** | -0.024** | -0.021** | -0.047*** | -0.072*** |
| | (0.003) | (0.010) | (0.010) | (0.014) | (0.017) |
| $\omega_{ik,t-1}^{emi}$ | 0.014* | 0.015* | 0.014 | -0.005 | -0.042 |
| | (0.008) | (0.009) | (0.009) | (0.017) | (0.054) |
| $\omega_{ik,t-1}^{births}$ | -0.003 | -0.004 | -0.003 | -0.037*** | -0.018 |
| | (0.009) | (0.010) | (0.011) | (0.008) | (0.032) |
| $diversity_{i,t-1}$ | -0.010 | -0.006 | 0.000 | | |
| | (0.015) | (0.032) | (0.038) | | |
| $ubiquity_{k,t-1}$ | -0.011 | -0.015** | -0.023*** | -0.021** | -0.052** |
| | (0.007) | (0.007) | (0.007) | (0.010) | (0.020) |
| $\rho_{ik,t-1}^{M}$ | 1.815** | 2.308** | 1.788* | 2.195** | 7.230*** |
| | (0.825) | (0.913) | (0.921) | (1.086) | (0.575) |
| $\rho_{ik,t-1}^{\omega}$ | -0.029 | -0.003 | 0.021 | 0.039 | 0.139 |
| | (0.018) | (0.038) | (0.043) | (0.054) | (0.161) |
| log ($pop_{i,t}$) | -0.030** | -0.058 | -0.062 | | |
| | (0.011) | (0.167) | (0.173) | | |
| FE: period | Y | Y | Y | | |
| FE: region | | Y | Y | | |
| FE: category | | | Y | Y | |
| FE: period-region | | | | Y | |
| FE: region-period-broad category | | | | | Y |
| FE: category-period | | | | | Y |
| Observations | 2070 | 2048 | 2042 | 2009 | 1056 |
| Pseudo-R2 | 0.111 | 0.129 | 0.136 | 0.148 | 0.225 |
| BIC | 2619.9 | 3475.0 | 3615.5 | 4040.7 | 3665.7 |

Standard errors are clustered by period and region. *asinh()* denotes the inverted hyperbolic sine function. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S33**. Logistic regression models explaining entries defined as first births in a specific activity, decomposing RCA values*

| | Dependent Variable: $Entry2_{ik,t}$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| asinh ($N_{ik,t-1}^{immi}$) | 0.203* | 0.195 | 0.256** | 0.317** | 0.325*** |
| | (0.117) | (0.136) | (0.118) | (0.143) | (0.114) |
| asinh ($\sum_k N_{ik,t-1}^{immi}$) | 0.051 | 0.338*** | 0.347** | | |
| | (0.056) | (0.095) | (0.117) | | |
| asinh ($\sum_i N_{ik,t-1}^{immi}$) | 1.704** | 1.706* | 0.918** | 0.809** | 3.208*** |
| | (0.842) | (0.912) | (0.316) | (0.374) | (1.103) |
| asinh ($\sum_k N_{ik,t-1}^{emi}$) | 0.024 | -0.481 | -0.582* | | |
| | (0.061) | (0.299) | (0.341) | | |
| asinh ($\sum_i N_{ik,t-1}^{emi}$) | -1.366 | -1.345 | -0.780 | -0.729 | -2.791 |
| | (0.866) | (0.990) | (0.482) | (0.522) | (1.640) |
| $\omega_{ik,t-1}^{immi}$ | 0.017*** | 0.015*** | 0.017*** | 0.023*** | 0.005 |
| | (0.002) | (0.004) | (0.004) | (0.006) | (0.006) |
| $\omega_{ik,t-1}^{emi}$ | 0.017*** | 0.012** | 0.012* | -0.004 | 0.012 |
| | (0.003) | (0.006) | (0.007) | (0.018) | (0.028) |
| $\omega_{ik,t-1}^{births}$ | -0.010 | -0.005 | -0.009 | 0.004 | -0.017 |
| | (0.007) | (0.007) | (0.007) | (0.010) | (0.015) |
| $diversity_{i,t-1}$ | 0.017*** | 0.015*** | 0.017*** | 0.023*** | 0.005 |
| | (0.018) | (0.048) | (0.049) | | |
| $ubiquity_{k,t-1}$ | -0.007 | -0.007 | -0.005 | -0.005 | -0.016 |
| | (0.005) | (0.006) | (0.009) | (0.010) | (0.014) |
| $\rho_{ik,t-1}^{M}$ | 1.704*** | 1.872*** | 2.651*** | 3.018*** | 2.988*** |
| | (0.326) | (0.355) | (0.527) | (0.698) | (0.520) |
| $\rho_{ik,t-1}^{\omega}$ | 0.053 | 0.047 | 0.085* | 0.098* | 0.187 |
| | (0.043) | (0.072) | (0.038) | (0.050) | (0.124) |
| log ($pop_{i,t}$) | 0.282*** | 0.383* | 0.419 | | |
| | (0.068) | (0.197) | (0.248) | | |
| FE: period | Y | Y | Y | | |
| FE: region | | Y | Y | | |
| FE: category | | | Y | Y | |
| FE: period-region | | | | Y | |
| FE: region-period-broad category | | | | | Y |
| FE: category-period | | | | | Y |
| Observations | 5600 | 5570 | 5570 | 5517 | 3541 |
| Pseudo-R2 | 0.125 | 0.154 | 0.231 | 0.245 | 0.325 |
| BIC | 5367.5 | 6292.7 | 6042.0 | 6558.2 | 8411.8 |

Standard errors are clustered by period and region. *asinh()* denotes the inverted hyperbolic sine function. * p < 0.1, ** p < 0.05, *** p < 0.01

***Table S34***. *Logistic regression models explaining exits defined as last births in a specific activity, decomposing RCA values*

| | Dependent Variable: $Exit2_{ik,t}$ | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| asinh ($N_{ik,t-1}^{immi}$ | -0.341*** | -0.360*** | -0.382*** | -0.410*** | -0.311** |
| | (0.053) | (0.067) | (0.084) | (0.076) | (0.120) |
| asinh ($\sum_k N_{ik,t-1}^{immi}$) | -0.250*** | -0.422** | -0.401** | | |
| | (0.047) | (0.168) | (0.193) | | |
| asinh ($\sum_i N_{ik,t-1}^{immi}$) | -1.640** | -1.498** | -1.259** | -1.574** | -14.026*** |
| | (0.661) | (0.584) | (0.527) | (0.544) | (3.014) |
| asinh ($N_{ik,t-1}^{emi}$) | -0.082 | -0.089 | -0.055 | -0.046 | 0.115 |
| | (0.064) | (0.063) | (0.091) | (0.105) | (0.177) |
| asinh ($\sum_k N_{ik,t-1}^{emi}$) | 0.053 | 0.151 | 0.109 | | |
| | (0.103) | (0.321) | (0.320) | | |
| asinh ($\sum_i N_{ik,t-1}^{emi}$) | 0.699 | 0.390 | 0.542 | 0.651 | 12.561** |
| | (0.797) | (0.686) | (0.585) | (0.707) | (3.796) |
| $\omega_{ik,t-1}^{immi}$ | 0.001 | -0.011** | -0.015*** | -0.016 | -0.064*** |
| | (0.008) | (0.005) | (0.005) | (0.014) | (0.024) |
| $\omega_{ik,t-1}^{emi}$ | 0.006 | 0.010 | 0.008 | 0.008 | 0.014 |
| | (0.010) | (0.009) | (0.008) | (0.016) | (0.012) |
| $\omega_{ik,t-1}^{births}$ | 0.002 | -0.014 | -0.013 | -0.023 | 0.005** |
| | (0.012) | (0.012) | (0.013) | (0.019) | (0.002) |
| $diversity_{i,t-1}$ | -0.044 | 0.071 | 0.076 | | |
| | (0.037) | (0.041) | (0.047) | | |
| $ubiquity_{k,t-1}$ | 0.012** | 0.012 | -0.008 | -0.002 | -0.005 |
| | (0.005) | (0.007) | (0.013) | (0.016) | (0.027) |
| $\rho_{ik,t-1}^{M}$ | -0.616 | -0.749 | -0.270 | -0.709 | -2.410 |
| | (1.058) | (0.939) | (0.929) | (0.870) | (1.892) |
| $\rho_{ik,t-1}^{\omega}$ | 0.053 | 0.128 | 0.094 | 0.112 | 0.127 |
| | (0.042) | (0.076) | (0.052) | (0.065) | (0.096) |
| log ($pop_{i,t}$) | -0.341*** | -0.058 | -0.118 | | |
| | (0.034) | (0.336) | (0.311) | | |
| FE: period | Y | Y | Y | | |
| FE: region | | Y | Y | | |
| FE: category | | | Y | Y | |
| FE: period-region | | | | Y | |
| FE: region-period-broad category | | | | | Y |
| FE: category-period | | | | | Y |
| Observations | 2687 | 2687 | 2659 | 2547 | 1498 |
| Pseudo-R2 | 0.209 | 0.263 | 0.327 | 0.332 | 0.405 |
| BIC | 3118.7 | 3947.3 | 3853.5 | 4278.8 | 4655.8 |

Standard errors are clustered by period and region. *asinh()* denotes the inverted hyperbolic sine function. * p < 0.1, ** p < 0.05, *** p < 0.01

# 4. References


Bairoch, P., Batou, J., & Chèvre, P. (1988). *La population des villes européennes de 800 à 1850*. Librairie Droz.

Buringh, E. (2021). The Population of European Cities from 700 to 2000: Social and Economic History. *Research Data Journal for the Humanities and Social Sciences*, *6*(1), 1–18. https://doi.org/10.1163/24523666-06010003

Elekes, Z., Boschma, R., & Lengyel, B. (2019). Foreign-owned firms as agents of structural change in regions. *Regional Studies*, *53*(11), 1603–1613. https://doi.org/10.1080/00343404.2019.1596254

Laouenan, M., Bhargava, P., Eyméoud, J.-B., Gergaud, O., Plique, G., & Wasmer, E. (2022). A cross-verified database of notable people, 3500BC-2018AD. *Scientific Data*, *9*(1), 290. https://doi.org/10.1038/s41597-022-01369-4

Miguelez, E., & Morrison, A. (2022). Migrant Inventors as Agents of Technological Change. *The Journal of Technology Transfer*. https://doi.org/10.1007/s10961-022-09927-z

Neffke, F., Hartog, M., Boschma, R., & Henning, M. (2018). Agents of Structural Change: The Role of Firms and Entrepreneurs in Regional Diversification. *Economic Geography*, *94*(1), 23–48. https://doi.org/10.1080/00130095.2017.1391691

Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., & Helbing, D. (2014). A network framework of cultural history. *Science*, *345*(6196), 558–562. https://doi.org/10.1126/science.1240064

Serafinelli, M., & Tabellini, G. (2022). Creativity over time and space: A historical analysis of European cities. *Journal of Economic Growth*, *27*(1), 1–43. https://doi.org/10.1007/s10887-021-09199-6

Yu, A. Z., Ronen, S., Hu, K., Lu, T., & Hidalgo, C. A. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data*, *3*(1), 150075. https://doi.org/10.1038/sdata.2015.75

# SUPPLEMENTARY MATERIALS
# FOR CHAPTER 3

# Content

# 1. Source data on GDP per capita levels

We compile several sources on GDP per capita levels:

Maddison project (1, 2) (2020 release) for country-level historical GDP per capita levels

Regional estimates of historical income levels for United Kingdom (3, 4), Sweden (5, 6), France (7, 8), Italy (9), Spain (10), Portugal (11) and Belgium (12) covering years prior to the 21st century. We match these estimates to NUTS2-regions (2021 classification).

Regional GDP per capita levels for the year 2000 from Eurostat (13), the Office for National Statistics in the UK (14), the Bureau of Economic Analysis in the United States (15), Statistics Canada (16), the State Statistics Service of Ukraine (17), Belstat in Belarus (18), and Rosstat in Russia (19).

We transform all data points to match 2011 USD PPP, matching the 2020 release of the Maddison project. In total, we obtain a dataset with 1,268 labeled observations in 50-year intervals (1300, 1350, …, 1950, 2000). All source data is reported in the published dataset.

**Adaptations.** We construct a dataset in 50-year intervals (1300, 1350, …, 1950, 2000). Not all observations in the historical datasets match these intervals. To increase our labeled dataset, we make slight adjustments to the source data in two forms: (1) If a source reports GDP per capita levels for e.g. 1545 and 1555, but not 1550 (Spain, for instance, in the Maddison project), we take the average GDP per capita level of 1545 and 1555 as observation in 1550. (2) If estimates in proximity (using ≤ 20 years as a rule of thumb) to a missing observation are available, we take the closest. For instance, the Maddison project provides an estimate for Belgium in 1812 but not in 1800. Then, we use the estimate for 1812 as estimate for the year 1800. Similarly, we take the regional GDP per capita estimates of 1968 provided by the Office for National Statistics (4) in the UK for 1950.

The following list describes such adaptations made to the Maddison project:

- BEL: value of 1812 used for 1800
- Value of 1820 used for 1800 for CAN, DNK, AUT, CSK, NOR, IRL
- Value of 1870 used for 1850 for BGR, HUN, IRL, ALB, CHE, ROU
- FRA: average of 1789 and 1820 used for 1800
- HRV: value of 1952 used for 1950
- IRL: value of 1913 for 1900

- SVN: value of 1952 used for 1950
- ITA: value of 1310 used for 1300
- EST: value of 1855 used for 1850

**Border changes.** Country borders have changed over the past centuries, which is also reflected in the Maddison project. The source materials of the Maddison project (10, 20–29) provide detailed information on which borders the respective estimates are referring to. For instance, data in the Maddison Project for Italy prior to the late 19th century refers only to Northern Italy. We take the following border changes into account when assigning biographies to geographies:

*Great Britain*: Data in the Maddison Project only refers to England prior to 1700. We, hence, treat England, Wales, Scotland and Northern Ireland as separate countries prior to 1700.

*Netherlands*: Data in the Maddison Project only refers to Holland (i.e. the NUTS-2 regions NL32 and NL33) prior to 1807.

*Italy*: Data in the Maddison Project only refers to Northern Italy (i.e. the NUTS regions ITC, ITH, ITI1, ITI2 and ITI3) prior to 1861.

*Germany*: Data in the Maddison project prior to 1850 refers to the "overlap between the Holy Roman Empire in the borders of 1792 and the territory of the nation state formed in 1871". Specifically, we take this into account by adding several regions of Poland (i.e. PL42, PL43, PL51, PL52, PL224, PL227, PL228, PL229, PL22B & PL22C) and Belgium (i.e. BE336) to Germany, while removing South Schleswig (i.e. DEF07 & DEF0C).

*Poland*: Data in the Maddison project prior to 1850 refers to the district of Cracow only.

*Czechoslovakia*: In the Maddison Project, estimates for Czechia or Slovakia do not exist prior to 1993, but just for Czechoslovakia (starting in 1820). Hence, we apply the borders of Czechoslovakia between 1820 and 1993. For earlier periods, however, we generate separate out-of-sample estimates for Czechia and Slovakia.

# 2. Geography

## 2.1. Regional classifications

We use the following geographical units for regions:

- European Union and EFTA countries: NUTS2-regions (Figure S11), 2021 edition
- Rest of Europe (BLR, UKR, RUS, BIH, MDA): oblasts or regions of similar size (Figure S11)

- United States: micropolitan and metropolitan statistical areas (Figure S12)
- Canada: Metropolitan areas (Figure S12)

The *.shp* files used in this study are provided in the replication package.

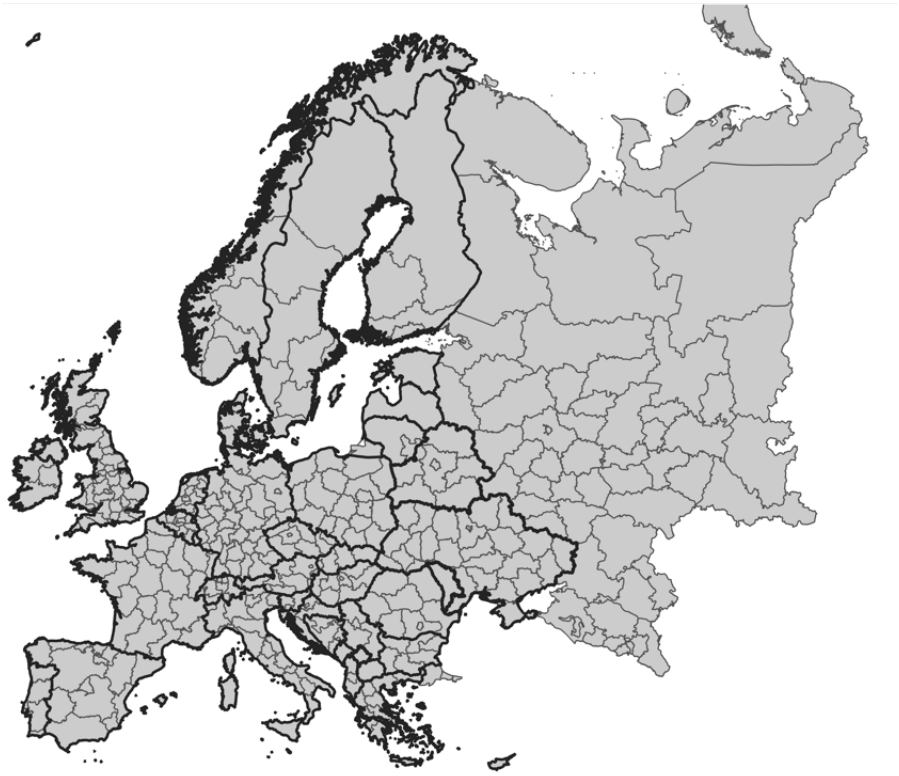


*Figure S11. Administrative borders in Europe (NUTS2 regions, oblasts and similar regions)*

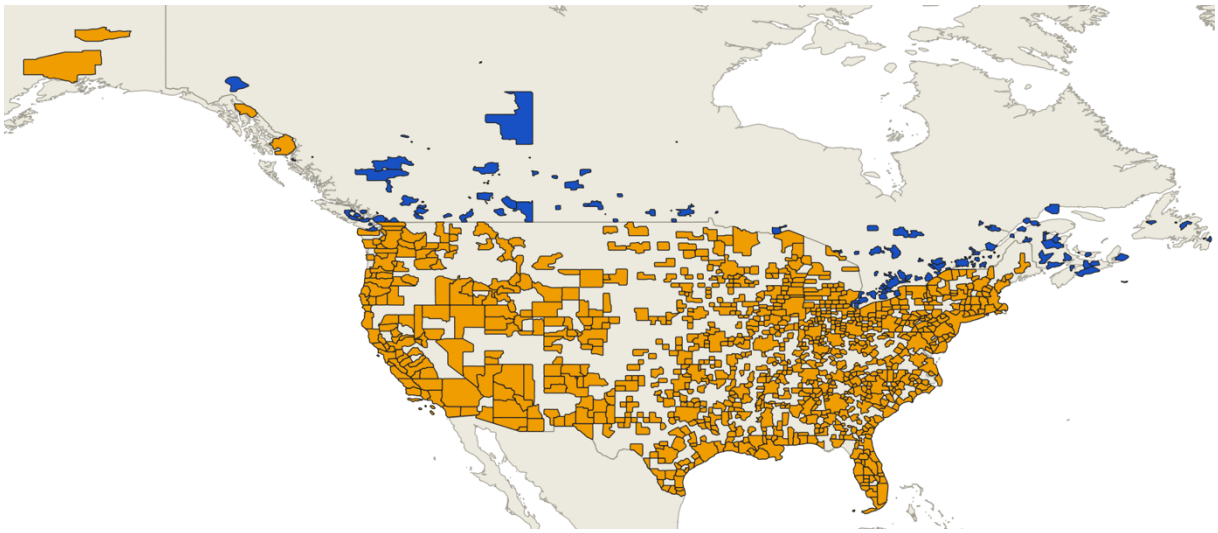


*Figure S12. Administrative borders in the United States (orange; micro- and metropolitan statistical areas) and Canada (blue; metropolitan areas)*

## 2.2. Supranational regions

We use supranational regions as fixed effects in our baseline model and potential candidates in the elastic net model. This classification mostly follows the UN geoscheme.

Specifically, we create dummy variables for the following region-time combinations. We need to aggregate early time periods due to limited observations:

- Southern Europe up to 1800
- Northern Europe up to 1800
- Western Europe before 1800
- Eastern Europe before 1800
- Northern America before 1800
- Western Europe in 1800
- Eastern Europe in 1800
- North America in 1800
- Southern Europe in 1850
- Northern Europe in 1850
- Western Europe in 1850
- Eastern Europe in 1850
- North America in 1850
- Southern Europe in 1900
- Northern Europe in 1900
- Western Europe in 1900
- Eastern Europe in 1900
- North America in 1900
- Southern Europe in 1950
- Northern Europe in 1950
- Western Europe in 1950
- Former Soviet Union in 1950
- North America in 1950
- Southern Europe in 2000
- Northern Europe in 2000
- Western Europe in 2000
- Former Soviet Union in 2000
- North America in 2000

One change we make to the UN geoscheme in assigning countries to supranational regions concern the Baltic states. We assign them to Northern Europe prior to 1750 (which they are originally in the UN geoscheme), to Eastern Europe between 1750 and 1950, and to former Soviet Union countries in 2000.

## 3. Data on famous individuals

We use a recently published and the most comprehensive database for notable people from Wikipedia, curated and cross-verified by Morgane Laouenan and colleagues (30). This database

collects data on 2.29 million famous individuals across human history, including their places of birth and death, their occupation, and proxies of their historical importance such as Wikipedia page views or the number of language editions.

## 3.1. Summary statistics

We assign biographies to countries and regions and use only those biographies that satisfy the following conditions:

- Wikipedia pages in at least two language editions
- An identifiable occupation

The latter is rooted in the fact that the granular occupation classification provided by Laouenan and colleagues is imperfect. Specifically, for the 633,820 famous individuals with at least two Wikipedia editions and living in Europe or the United States between 1150 and 2000, the database shows 2,750 unique occupations, differentiating between e.g. *actor* and *actress*, *designer* and *fashion designer* or *zoologist* and *biologist*. We manually clean the occupations and derive a classification with 49 unique occupations.

Figure S13 provides a treemap of the distribution across occupations in the dataset.

Table S35 shows the unbalanced distribution of biographies across time. While the dataset includes 1.417 individuals born between 1150 and 1299, it provides information on 364.252 individuals born between 1850 and 1999.

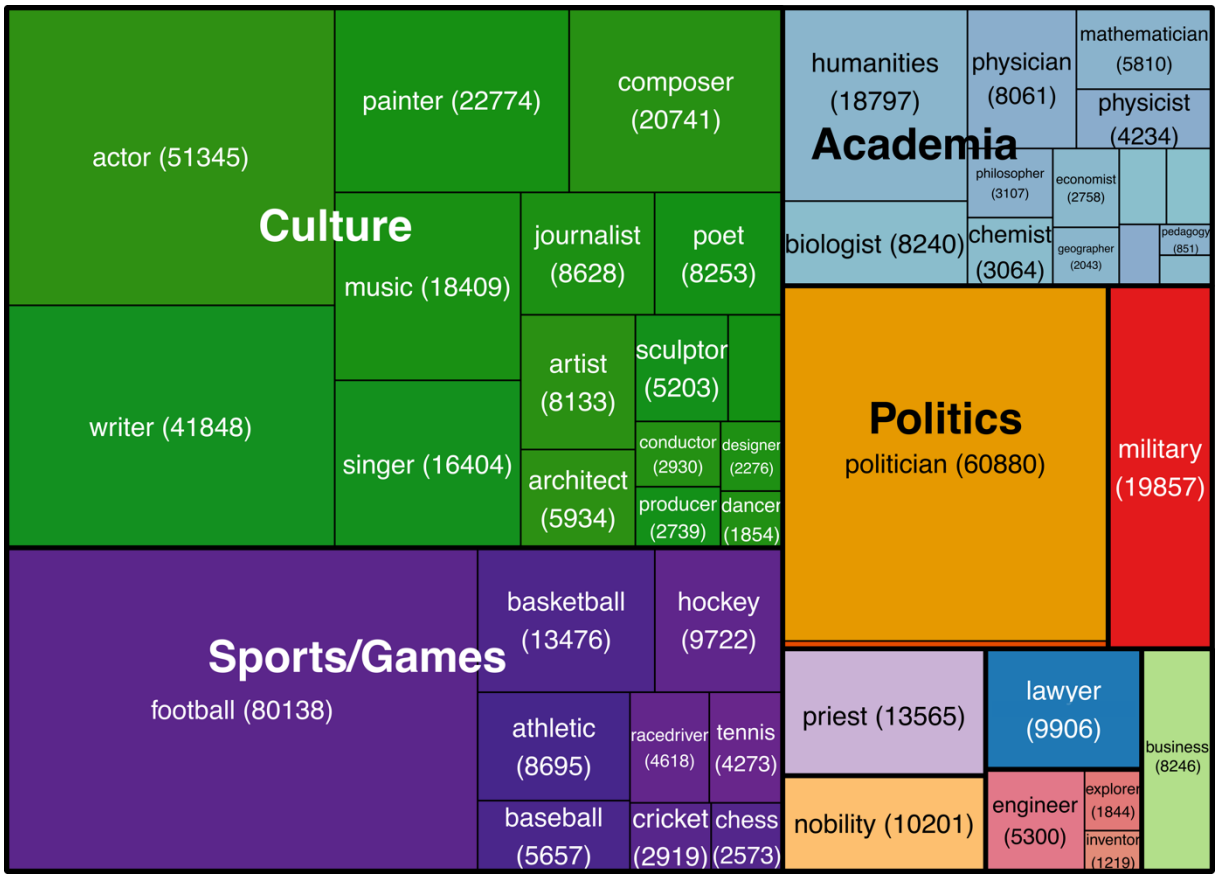



*Figure S13. Treemap of occupations in the dataset on famous individuals.*

*Table S35. Number of famous individuals across time periods*

| Period | Born after | Born before | No. of observations |
|---|---|---|---|
| 1 | 1150 | 1299 | 1.417 |
| 2 | 1200 | 1349 | 1.664 |
| 3 | 1250 | 1399 | 1.965 |
| 4 | 1300 | 1449 | 2.676 |
| 5 | 1350 | 1499 | 4.525 |
| 6 | 1400 | 1549 | 7.005 |
| 7 | 1450 | 1599 | 8.960 |
| 8 | 1500 | 1649 | 10.653 |
| 9 | 1550 | 1699 | 11.659 |
| 10 | 1600 | 1749 | 15.376 |
| 11 | 1650 | 1799 | 29.147 |
| 12 | 1700 | 1849 | 60.258 |
| 13 | 1750 | 1899 | 123.974 |
| 14 | 1800 | 1949 | 241.214 |
| 15 | 1850 | 1999 | 364.252 |

## 3.2. Historical Popularity Index (HPI)

We take the historical popularity of the individuals in our dataset into account. We follow the Historical Popularity Index (HPI) which has been introduced in the Pantheon database (31). In its original version, the HPI takes the individual's age, the number of language editions, the effective number of language editions based on the entropy in terms of page views across languages, the page views in non-English Wikipedia editions, and the variance of page views across different language editions into account.

We reconstruct the HPI with information we have available in the dataset by Laouenan and colleagues (30). Specifically, an individuals' $HPI$ is proportional to the number of Wikipedia page views ($V$), the number of language editions ($L$) and age ($A$, i.e. 2023 minus year of birth):

$$HPI = \begin{cases} log_{10}(V) + ln(L) + log_4(A) & if\ A \geq 70 \\ log_{10}(V) + ln(L) + log_4(A) - \dfrac{70 - A}{7} & if\ A < 70 \end{cases}$$

To assess how similar this measure of historical importance is to the HPI in the Pantheon database, we correlate these two values for the subset of famous individuals who are present in both datasets. We find that our measure of historical importance is highly correlated with the HPI in the Pantheon dataset ($R^2 = 0.76$, see Figure S14).

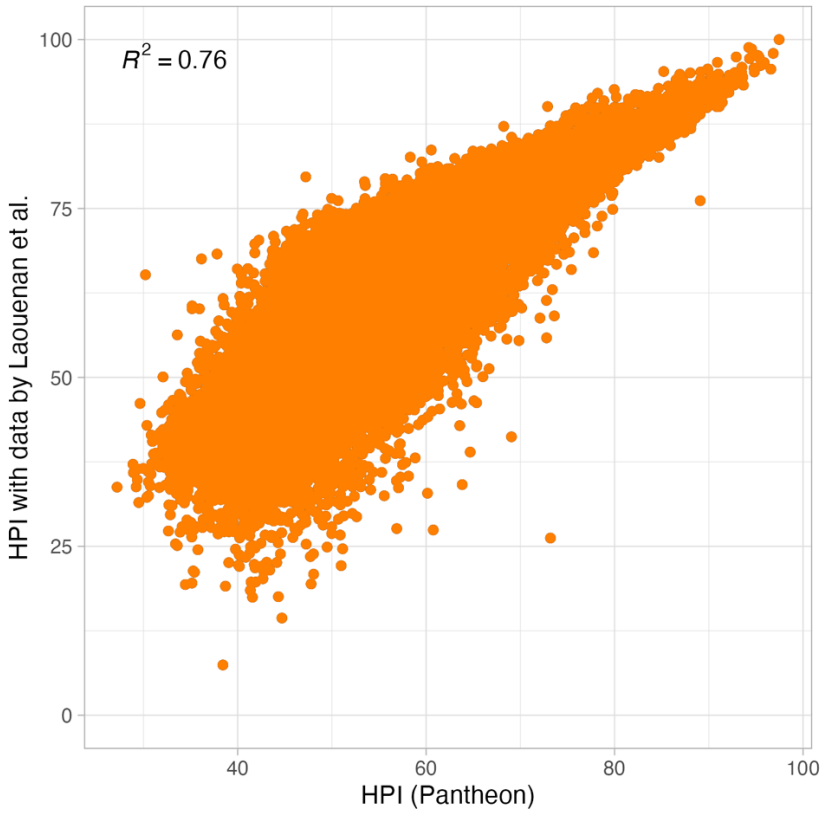



*Figure S14. Correlation between the Historical Popularity Index in the Pantheon dataset and the dataset curated by Laouenan et al.*

## 3.3. Famous individuals as proxy for population levels

Historical population data describes solely urban population (32, 33). To obtain population-weighted distributions within countries, we want to have data on population levels in rural regions as well. Hence, we use the number of famous births and deaths in a location as proxy for population. Figure S15 shows the correlation between existing data on urban population (32, 33) and the number of births and deaths in our dataset.

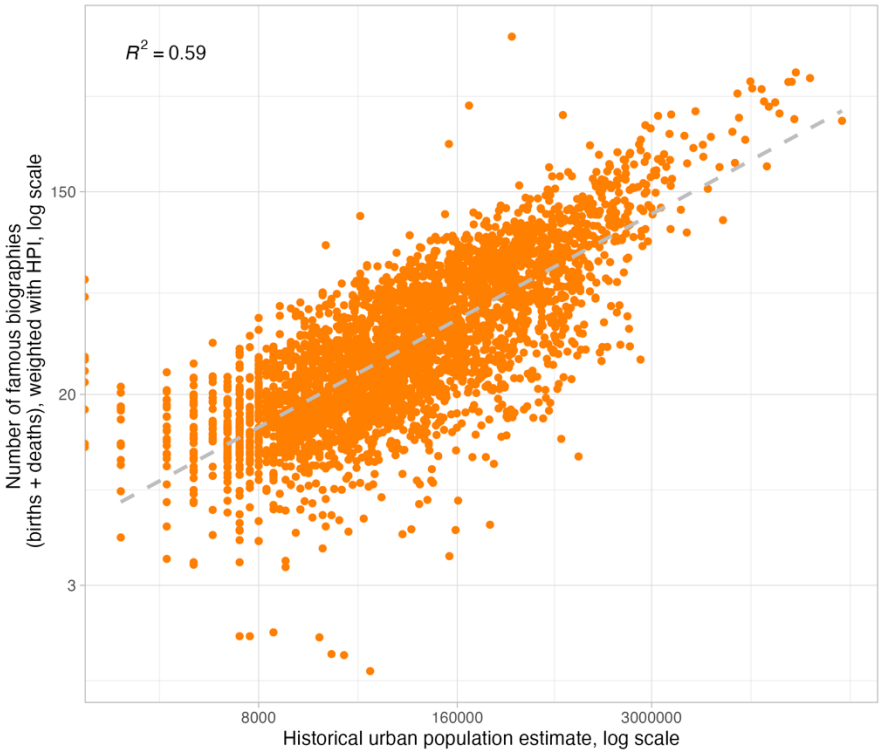



*Figure S15. Correlation between historical population data and number of famous biographies in a location*

## 3.4. Migration

We use places of birth and death as a proxy for migration, following the literature using similar data to describe migration movements (34, 35).

This evokes the question whether this proxy is valid. In a recent publication (36) we explored this question by randomly drawing ~200 individuals from the dataset, paying attention to representativeness across centuries. We read the Wikipedia article for each famous individual to determine whether a relation to the place of death exists, which would qualify as migration. We differentiated between (a) having any relation to the place of death (i.e. living there for a considerable amount of time, having noteworthy social connections with multiple visits there,

or, in case of politicians and noblemen, reigning over the region) and (b) having a major relation to the place of death. The latter is the case if the place was one of the individual's main places of living, if the famous individual taught at a university there etc.

We found that in 181 out of 202 cases ($\hat{p} = 0.896$, 95% CI: [0.854, 0.938]), the famous individual had a relation to his or her place of death. Hence, only in 10% of observations the place of death is arbitrary. Also, we found that in 151 out of 202 cases ($\hat{p} = 0.748$, 95% CI: [0.688, 0.807]), the famous individual had a major relation to his or her place of death. These results indicate that using place of birth and death as a proxy for migration is a valid approach. The sampled data is available in the GitHub repository associated with this publication (folder *misc/migration_proxy*).

# 4. Methods

## 4.1. Economic Complexity

We compute economic complexity indices for famous births, deaths, immigrants, and emigrants in a location to include them as potential features in our elastic net model. Here, we provide tables showing the 30 most complex locations in 1300 (Figure S16), 1600 (Figure S17), and 1900 (Figure S18) when considering famous births, deaths, immigrants, or emigrants.

A crucial methodological step in calculating the Economic Complexity Index is to make sure we compare locations and occupations that are not too different with respect to size. Hence, we do not compute ECI values for locations with less than eight births or deaths in a period up to 1600, with less than 20 births or deaths per period between 1650 and 1950, or with less than 50 births or deaths in 2000. For locations with fewer famous individuals, we impute the minimum ECI value.

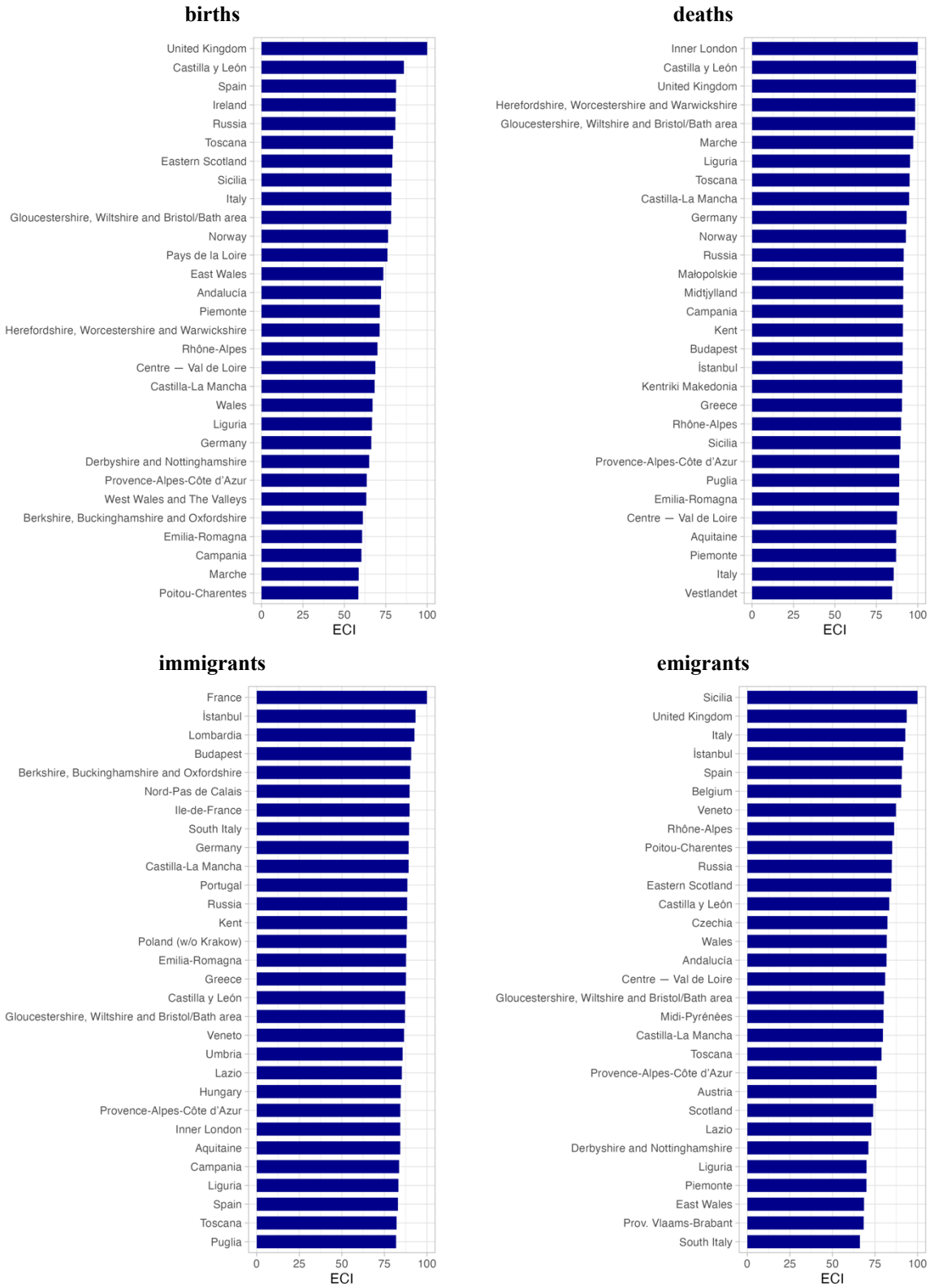



*Figure S16. The 30 locations with the highest Economic Complexity Index in 1300 for births, deaths, immigrants, and emigrants.*

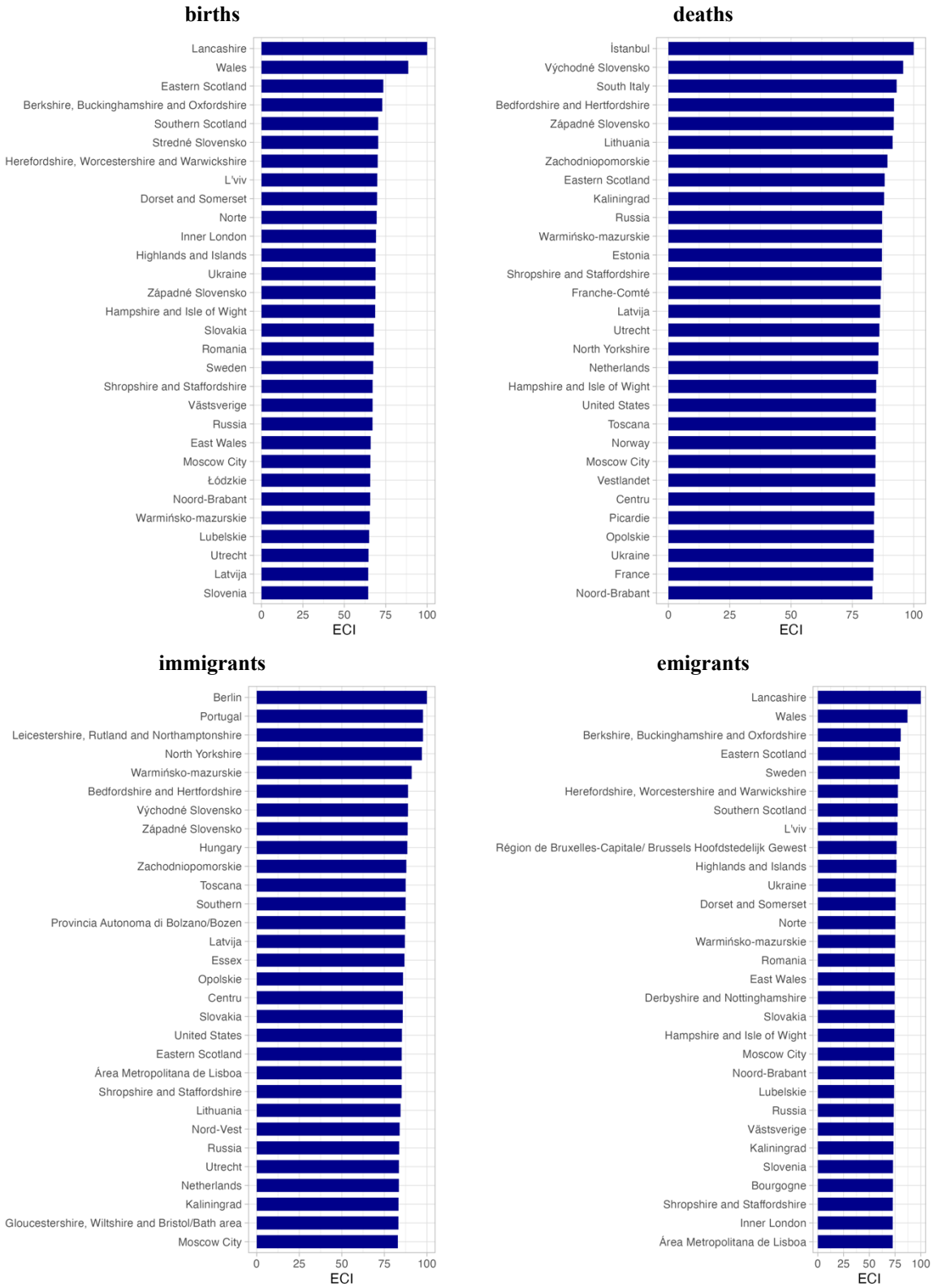



*Figure S17. The 30 locations with the highest Economic Complexity Index in 1600 for births, deaths, immigrants, and emigrants.*

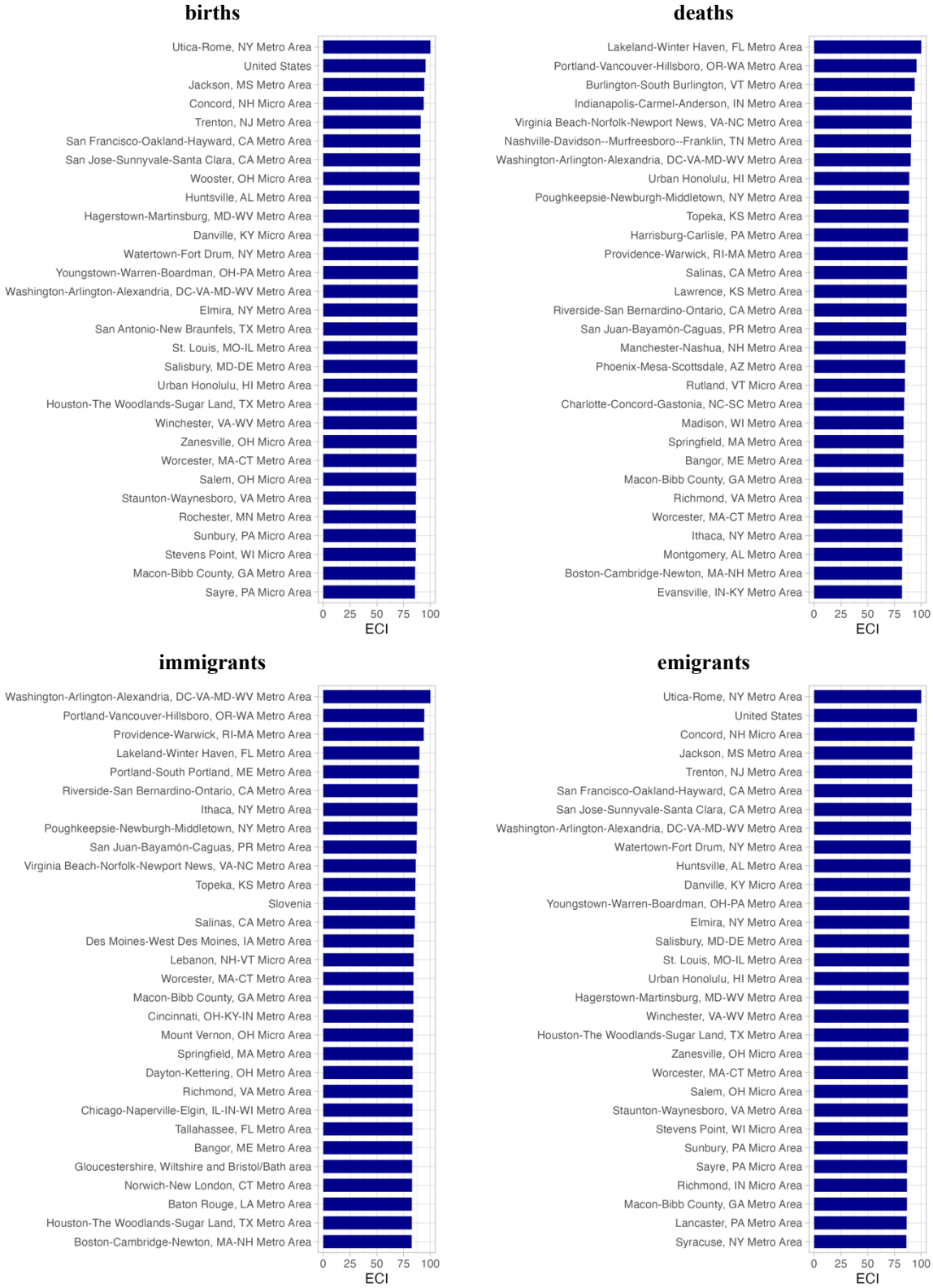



*Figure S18. The 30 locations with the highest Economic Complexity Index in 1900 for births, deaths, immigrants, and emigrants.*

## 4.2. Singular Value Decomposition

Singular Value Decomposition (SVD) is a dimensionality reduction technique which retrieves factors from rectangular matrices that best explain the structure of the underlying matrix. SVD is a generalization of the eigenvalue decomposition.

An overview of SVD and its connection to Cobb-Douglas production functions can be found in a recent review on economic complexity (37).

Two SVD factors that are selected by the elastic net model and play a significant role according to the Shapley values are the third factor of $N_{ik,1600}^{births}$ and the fourth factor of $N_{ik,1600}^{deaths}$. Figure S19 and Figure S20 plot these factors in a scatterplot with a measure of size on the vertical axis (total number of births in the location). Interpreting these factors is non-trivial. The third SVD factor of $N_{ik,1600}^{births}$ (Figure S19), for instance, seems to distinguish between some Dutch, and British regions on one side and Italian, French, and German regions on the other.

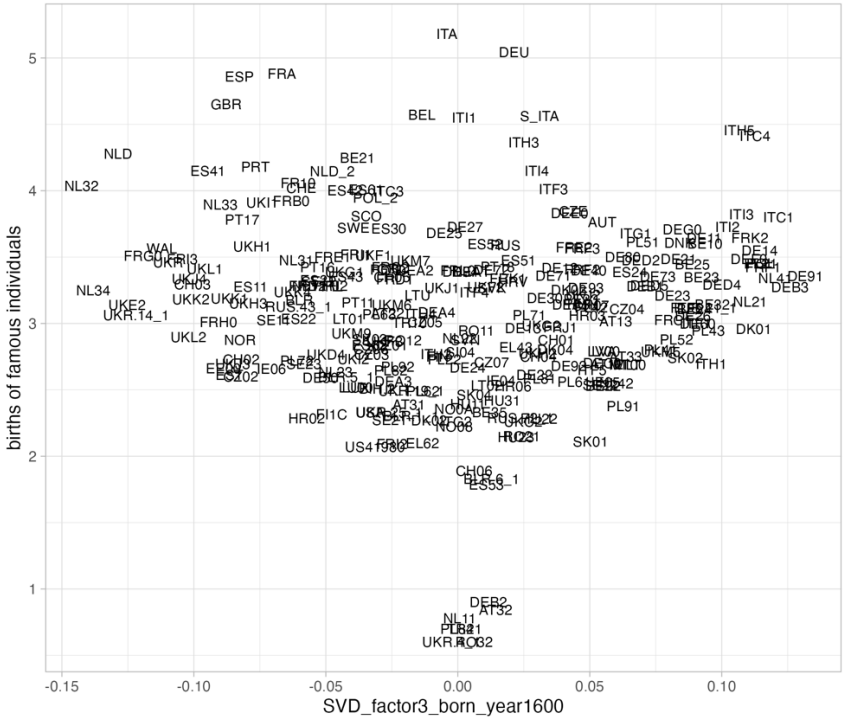



*Figure S19. Third SVD factor for births, 1600*

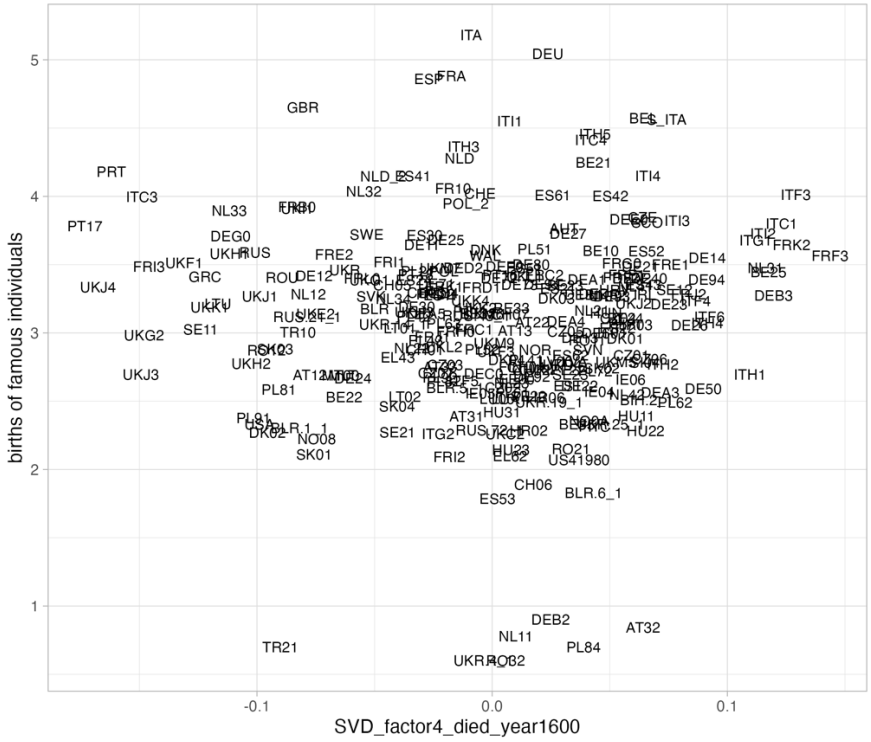


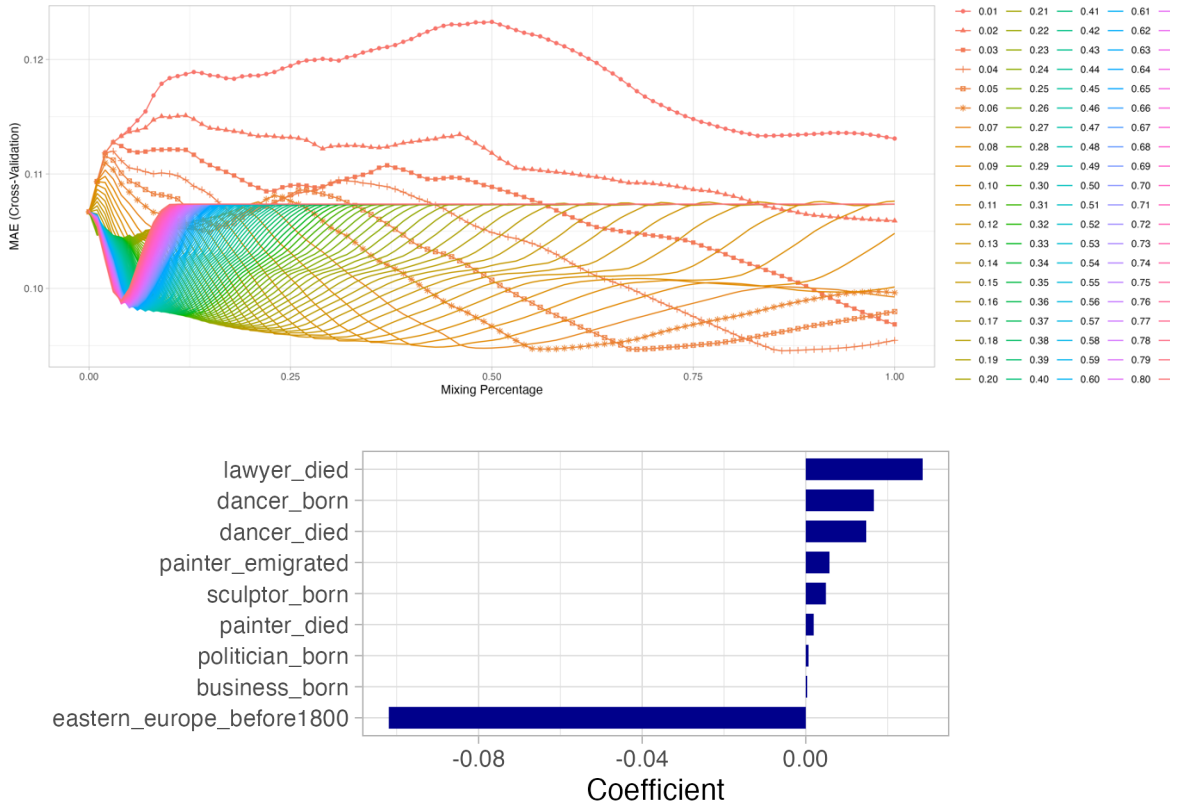

*Figure S20. Fourth SVD factor for deaths, 1600*

# 5. Results

## 5.1. EN model results

Here, we provide the results from the EN model for each period. This includes the plots showing the cross-validation of the model parameters as well as the selected features and their coefficients. Regularization models such as EN do not provide standard errors directly, which is why we use bootstrapping to obtain confidence intervals for our estimates. We aggregated our data to five historically informed epochs, i.e. the Late Middle Ages (1300-1500, Figure S21), the Early Modern Period (1550-1750, Figure S22), the Industrial Revolution (1800-1850, Figure S23), the Machine Age (1900-1950, Figure S24), and the Information Age (2000, Figure S25).

Also, we investigate whether the full model provides more explanatory power with respect to within-period variation, compared to the baseline model. Indeed, this seems to be the case, especially in early time periods, but the differences are relatively small (Figure S26).

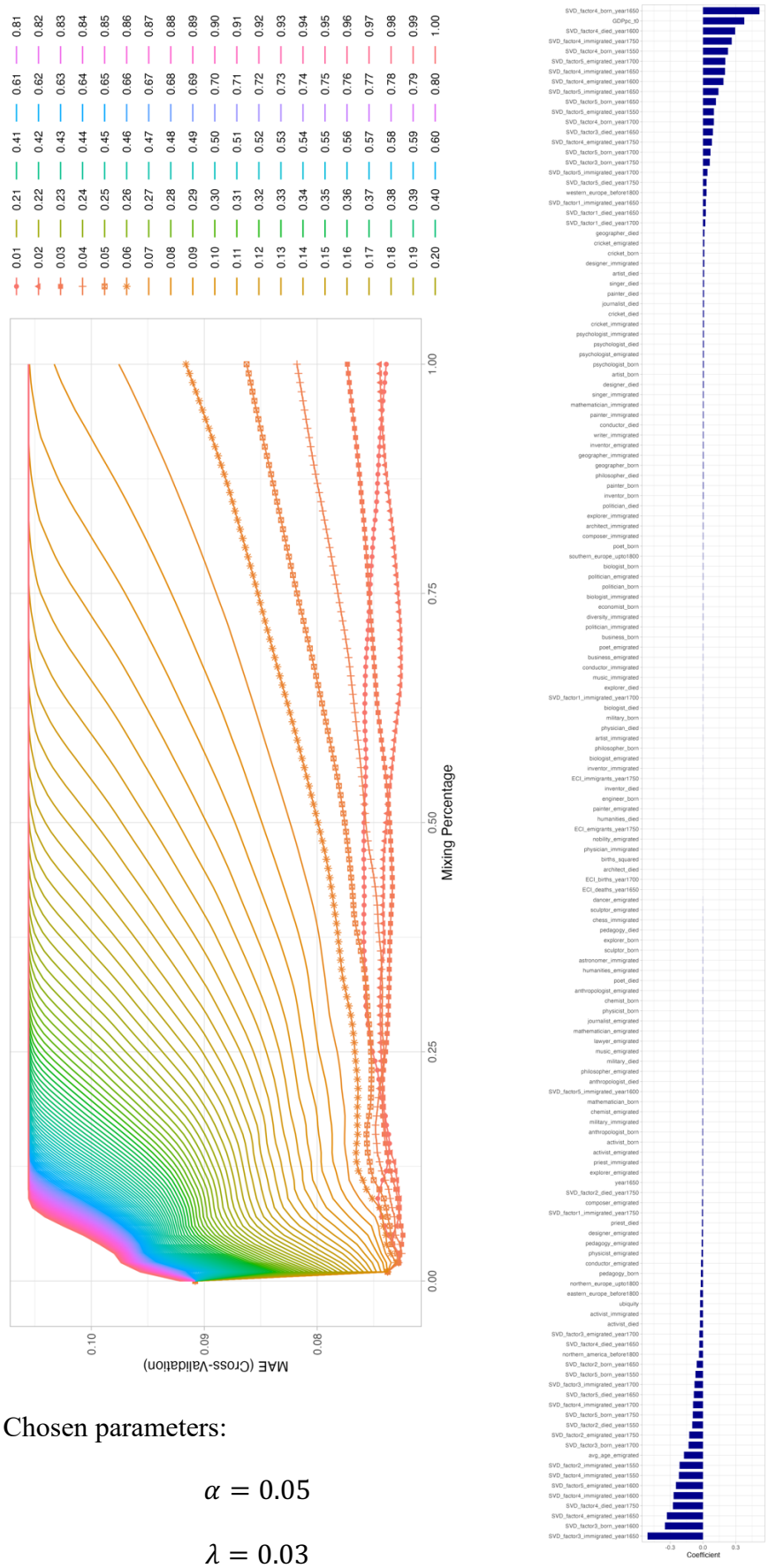



Chosen parameters:

$$\alpha = 0.86$$

$$\lambda = 0.04$$

*Figure S21. Model results for Late Middle Ages (1300-1500).*

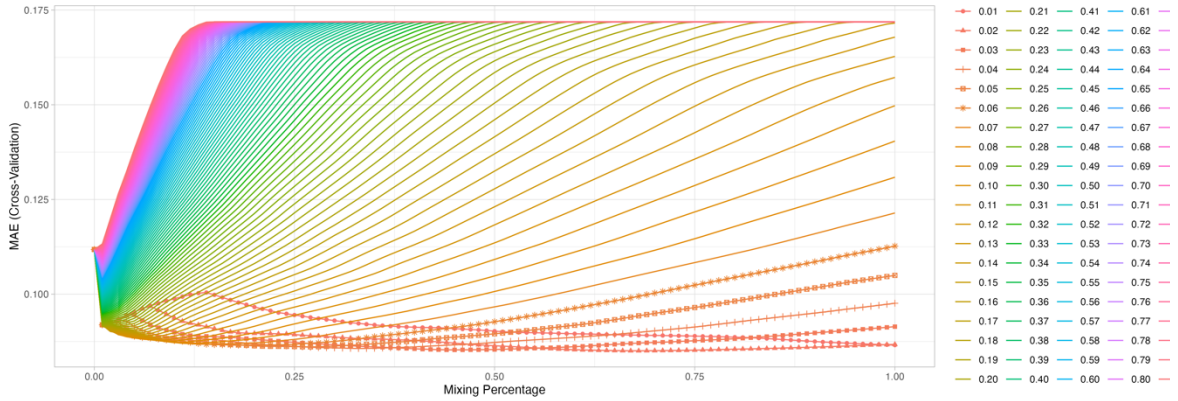


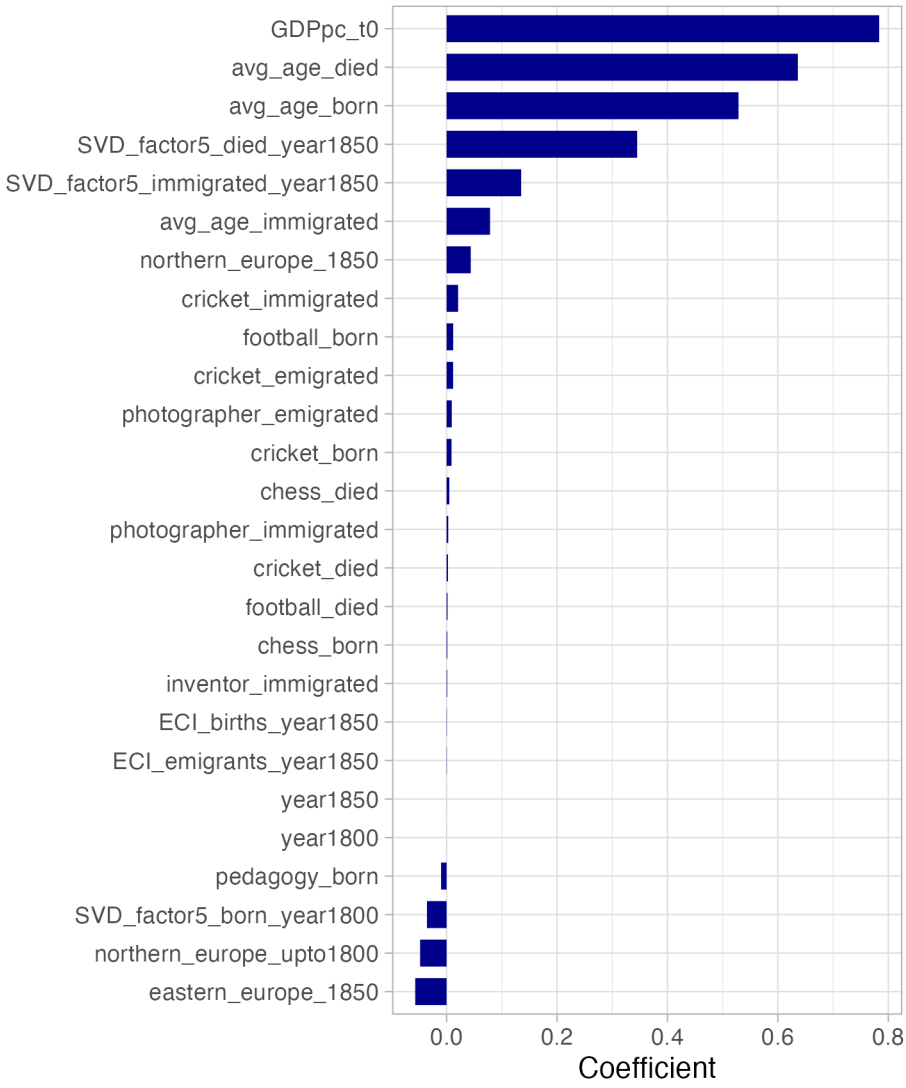

Chosen parameters:

$$\alpha = 0.05$$

$$\lambda = 0.03$$

*Figure S22. Model results for Early Modern Period (1550-1750).*



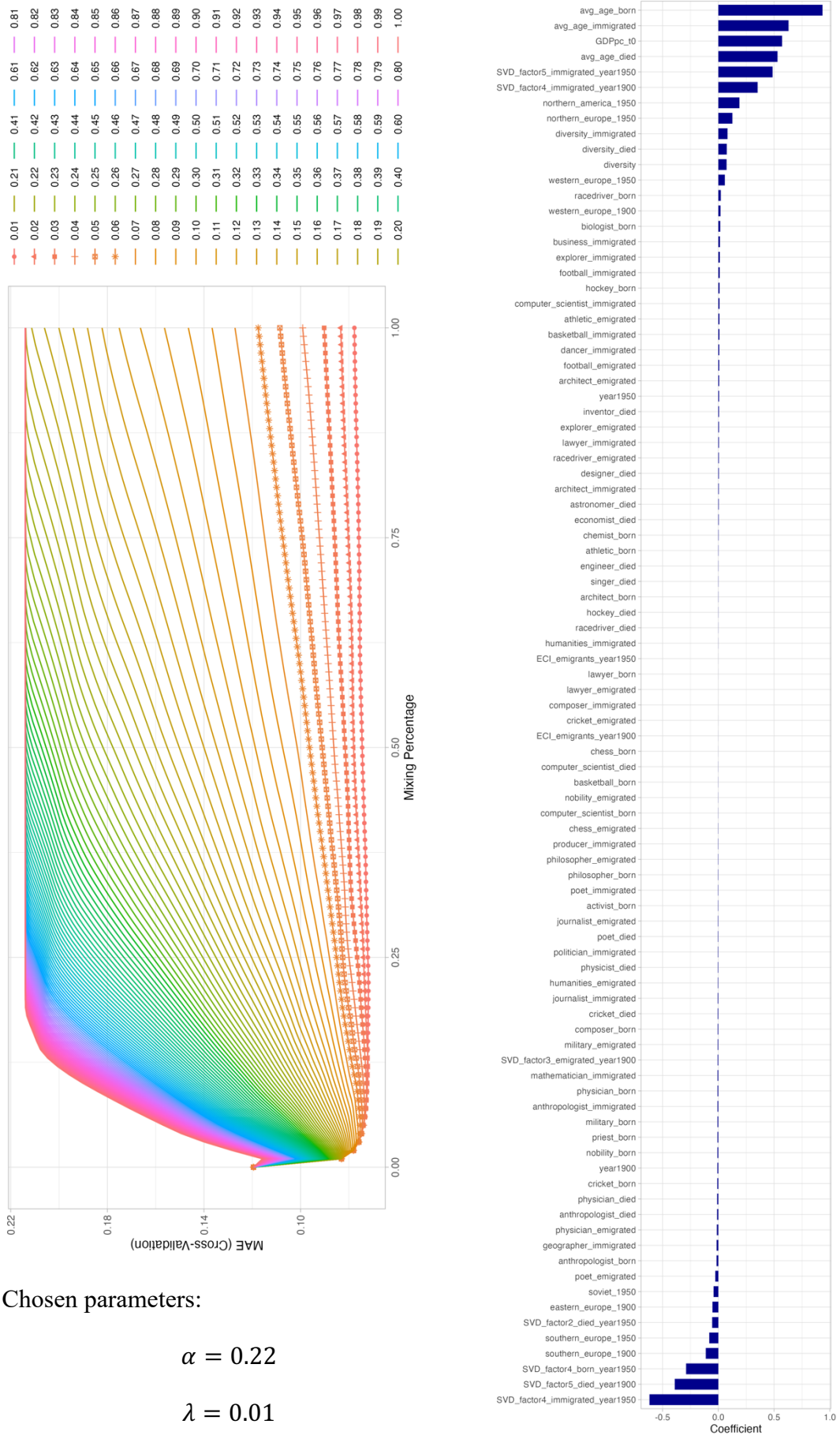




Chosen parameters:

$$\alpha = 0.68$$

$$\lambda = 0.02$$

*Figure S23. Model results for Industrial Revolution (1800-1850).*

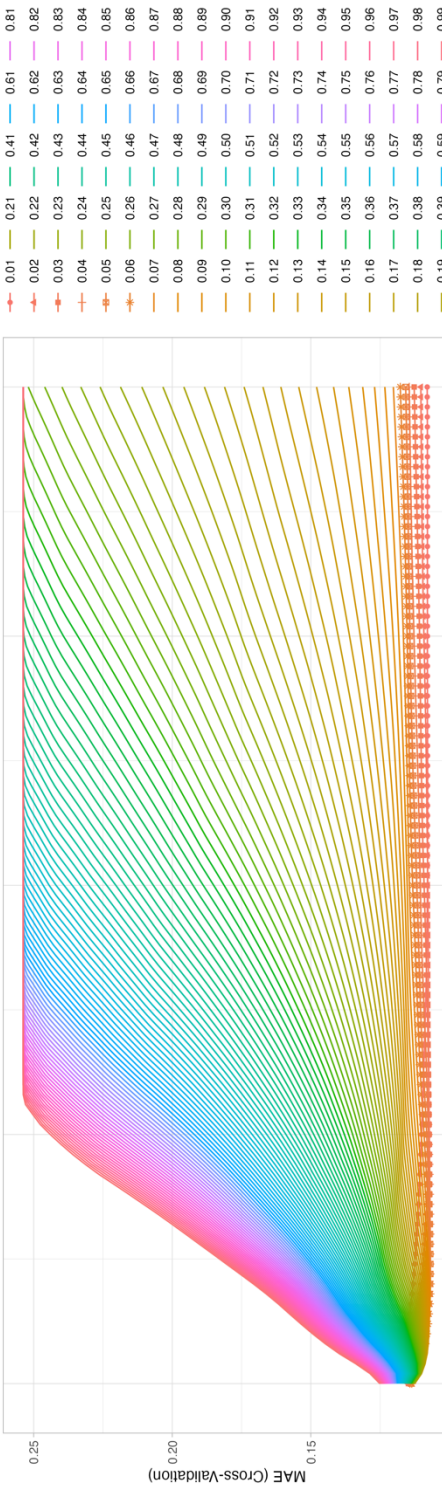



Chosen parameters:

$$\alpha = 0.22$$

$$\lambda = 0.01$$

*Figure S24. Model results for Machine Age (1900-1950).*





Chosen parameters:

$$\alpha = 0.1$$

$$\lambda = 0.06$$

*Figure S25. Model results for year 2000.*

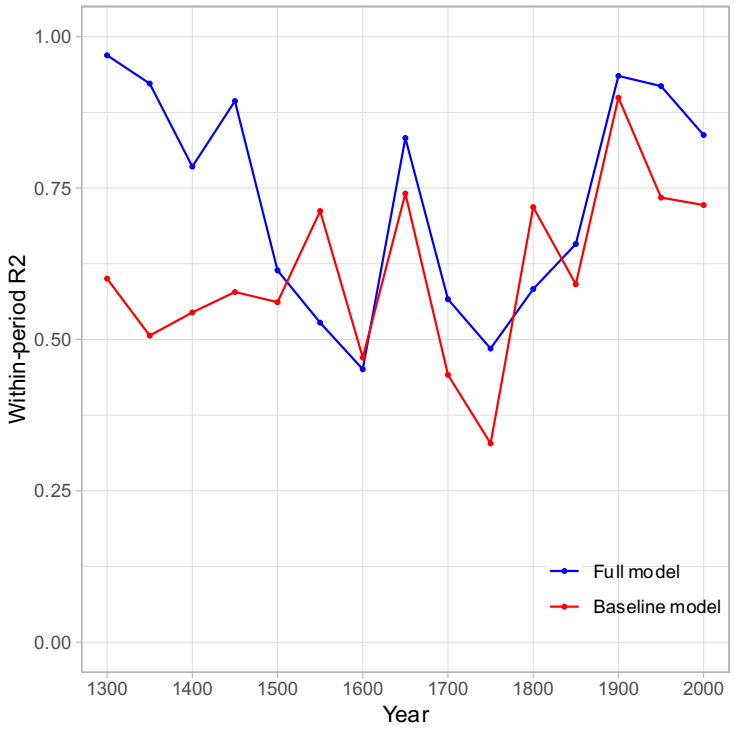



*Figure S26. Explanatory power of the models within-period.*

## 5.2. Atlantic trade

We follow the study by Acemoglu, Johnson & Robinson (38) to investigate the role of Atlantic trade in explaining the divergence between the North (UK, NLD, BEL) and the South (ITA, ESP, PRT) of Europe.

We use the information the authors provide in their paper, the appendix, and in the published data to recreate the subsets of cities and regions they identify having Atlantic and Mediterranean ports. Specifically, these are the NUTS-2 regions with Atlantic and Mediterranean ports:

- **Atlantic**: UKK4, UKI1, UKI2, UKK1, UKM8, NL32, NL33, NL34, FRD2, FRI2, ES61, ES13, PT11, PT17
- **Mediterranean**: ES62, ES52, ES51, FRJ1, FRL0, HR03, EL61, EL63, ITC2, ITC3, ITF1, ITF3, ITF4, ITG1, ITG2, ITH3, ITH4, ITH5, ITI1, ITI3, ITI4

We use the population-weighted average to aggregate these groups, using the number of births and deaths of famous individuals as a proxy for population.

Our results strongly resemble the results Acemoglu et al. present in their Fig. 2 for country-level development and Figs. 4 and 5 using city population as a proxy for regional economic development.

## 5.3. Generalizability of the results

We test the generalizability of our results in two ways.

First, we evaluate whether the labeled training data is different from unlabeled data with respect to some general features. Table S36 and Table S37 provide descriptive statistics of labeled and unlabeled observations for the GDP per capita levels in 2000, and the number of famous births, deaths, immigrants, and emigrants for countries and regions, respectively. Each observation in these tables refers to a country-year or region-year combination.

Second, we investigate the correlations between our estimated GDP per capita levels and proxies of economic development differentiated by labeled and unlabeled observations. Despite the differences in the descriptive statistics, we find encouraging results. That is, the correlations are highly similar for labeled and unlabeled observations (Figure S27, Figure S28, and Figure S29).

*Table S36. Descriptive statistics for labeled and unlabeled country-level observations*

| Variable | Period | labeled observations | | | unlabeled observations | | |
|---|---|---|---|---|---|---|---|
| | | *N (country-year)* | *mean* | *sd* | *N (country-year)* | *mean* | *sd* |
| **GDPpc [2000]** | 1300-1500 | 34 | 30977.6 | 6476.8 | 86 | 23253.6 | 14454.6 |
| | 1550-1750 | 55 | 30745 | 7312.8 | 106 | 22172.4 | 14346.8 |
| | 1800-1850 | 41 | 30791.5 | 11209.3 | 32 | 15575.6 | 13083 |
| | 1900-1950 | 52 | 29025.1 | 12996.3 | 25 | 11309 | 9453.8 |
| | 2000 | | | | | | |
| **births** | 1300-1500 | 34 | 254.2 | 319.8 | 114 | 33.9 | 51.2 |
| | 1550-1750 | 55 | 927.9 | 985 | 129 | 119.8 | 189 |
| | 1800-1850 | 42 | 2064.4 | 3024 | 38 | 327.5 | 480.9 |
| | 1900-1950 | 54 | 7045.1 | 10508.6 | 29 | 1021.7 | 1960.1 |
| | 2000 | 41 | 10858.7 | 18404.5 | 2 | 321.5 | 340.1 |
| **deaths** | 1300-1500 | 34 | 187.9 | 235.4 | 114 | 32.5 | 60.6 |
| | 1550-1750 | 55 | 722.9 | 793.6 | 129 | 126 | 247.8 |
| | 1800-1850 | 42 | 1790.1 | 2790.1 | 38 | 311.9 | 536 |
| | 1900-1950 | 54 | 5373.1 | 8789.4 | 29 | 565.4 | 1568.7 |
| | 2000 | 41 | 3885.5 | 7855.5 | 2 | 34.5 | 17.7 |
| **immigrants** | 1300-1500 | 34 | 100.7 | 101.1 | 114 | 22.5 | 43.1 |
| | 1550-1750 | 55 | 412.3 | 478 | 129 | 88.2 | 166.8 |
| | 1800-1850 | 42 | 1190.5 | 1968.1 | 38 | 208.5 | 380.2 |
| | 1900-1950 | 54 | 3886.9 | 7150.8 | 29 | 402.5 | 1253.4 |
| | 2000 | 41 | 2845.7 | 6496.7 | 2 | 13.5 | 0.7 |
| **emigrants** | 1300-1500 | 34 | 167.2 | 177.2 | 114 | 23.9 | 33 |
| | 1550-1750 | 55 | 617.5 | 648 | 129 | 82 | 111.5 |
| | 1800-1850 | 42 | 1465 | 2195.7 | 38 | 223.9 | 330.3 |
| | 1900-1950 | 54 | 5558.8 | 8888.4 | 29 | 859 | 1663.1 |
| | 2000 | 41 | 9818.9 | 17046.3 | 2 | 300.5 | 321.7 |

*Table S37. Descriptive statistics for labeled and unlabeled regional observations*

| Variable | Period | labeled observations | | | unlabeled observations | | |
|---|---|---|---|---|---|---|---|
| | | *N (region-year)* | *mean* | *sd* | *N (region-year)* | *mean* | *sd* |
| **GDPpc [2000]** | 1300-1500 | 0 | NaN | | 656 | 28200.5 | 10903.3 |
| | 1550-1750 | 43 | 25890.9 | 6727.9 | 1083 | 27094.9 | 12788.3 |
| | 1800-1850 | 111 | 29028.3 | 8602.1 | 687 | 28566.1 | 15970.2 |
| | 1900-1950 | 166 | 30714.7 | 9985.2 | 1179 | 30762.8 | 16196.8 |
| | 2000 | | | | | | |
| **births** | 1300-1500 | 0 | NaN | | 656 | 17.7 | 32.5 |
| | 1550-1750 | 43 | 78.8 | 69.6 | 1093 | 56 | 91 |
| | 1800-1850 | 111 | 249.1 | 376.5 | 767 | 89 | 140 |
| | 1900-1950 | 166 | 529.2 | 657.9 | 1488 | 207.3 | 501.8 |
| | 2000 | 790 | 537.6 | 1050.9 | 48 | 169.7 | 281.9 |
| **deaths** | 1300-1500 | 0 | NaN | | 656 | 18.1 | 29.6 |
| | 1550-1750 | 43 | 65.2 | 78.8 | 1093 | 52.2 | 126.6 |
| | 1800-1850 | 111 | 228.4 | 717.6 | 767 | 79.8 | 217 |
| | 1900-1950 | 166 | 374.9 | 701.1 | 1488 | 166 | 631.1 |
| | 2000 | 790 | 201.2 | 647.8 | 48 | 40.9 | 51.2 |
| **immigrants** | 1300-1500 | 0 | NaN | | 656 | 12 | 17.6 |
| | 1550-1750 | 43 | 44.1 | 61.1 | 1093 | 33 | 82.6 |
| | 1800-1850 | 111 | 147.4 | 494.9 | 767 | 53.6 | 152.9 |
| | 1900-1950 | 166 | 251 | 522.5 | 1488 | 122.5 | 478.5 |
| | 2000 | 790 | 147.7 | 505.3 | 48 | 33.6 | 42.5 |
| **emigrants** | 1300-1500 | 0 | NaN | | 656 | 11.6 | 16 |
| | 1550-1750 | 43 | 57.7 | 51.6 | 1093 | 36.8 | 43.9 |
| | 1800-1850 | 111 | 168.1 | 173.9 | 767 | 62.9 | 81.2 |
| | 1900-1950 | 166 | 405.4 | 487.1 | 1488 | 163.9 | 356.6 |
| | 2000 | 790 | 484.1 | 902.3 | 48 | 162.4 | 270.1 |

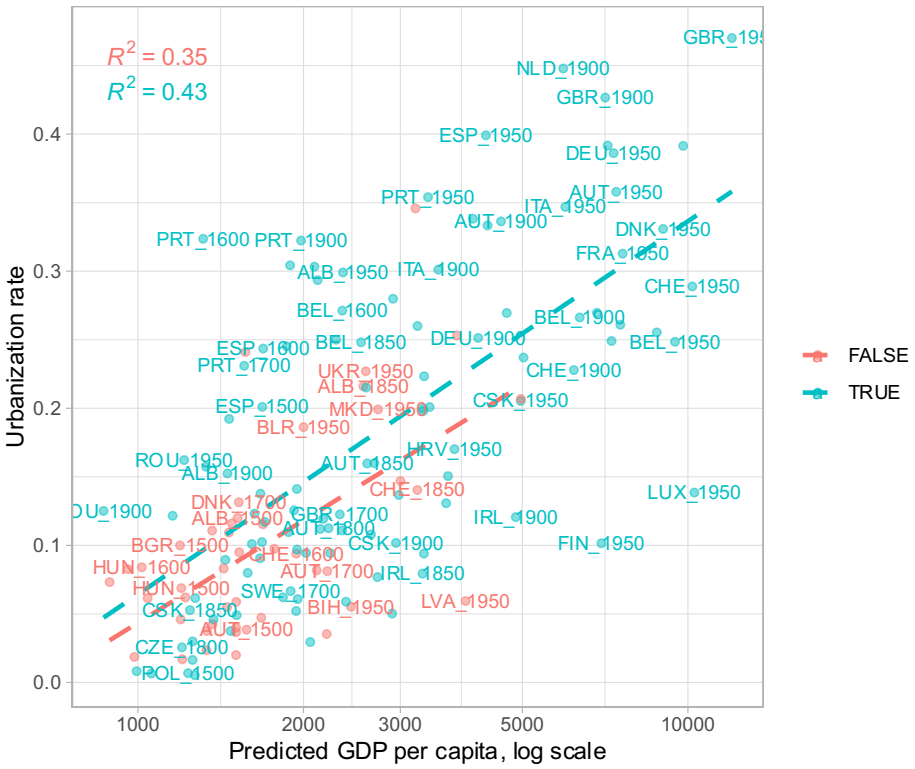



*Figure S27. Correlation between estimated GDP per capita and urbanization, for labeled (TRUE) and unlabeled (FALSE) observations*

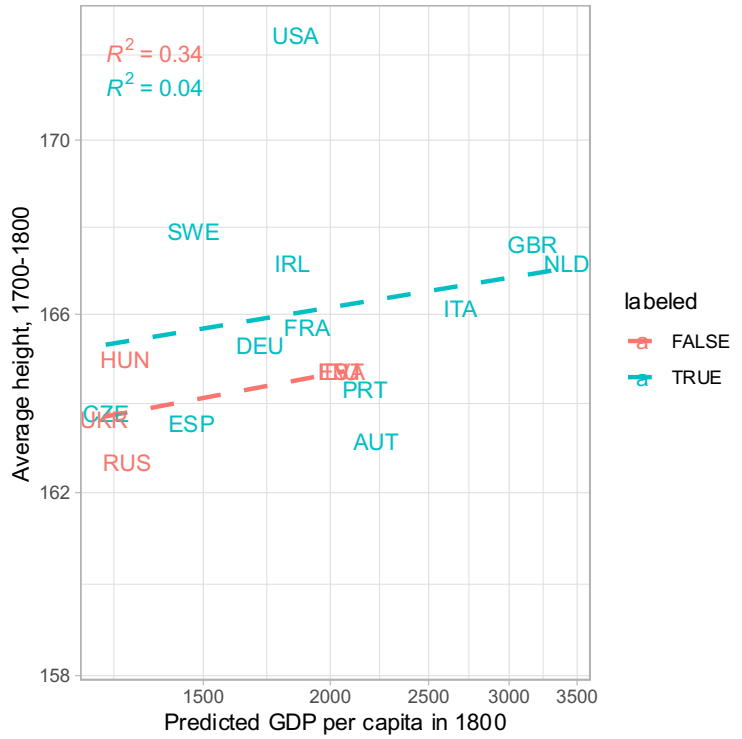



*Figure S28. Correlation between estimated GDP per capita and average body height, for labeled (TRUE) and unlabeled (FALSE) observations*

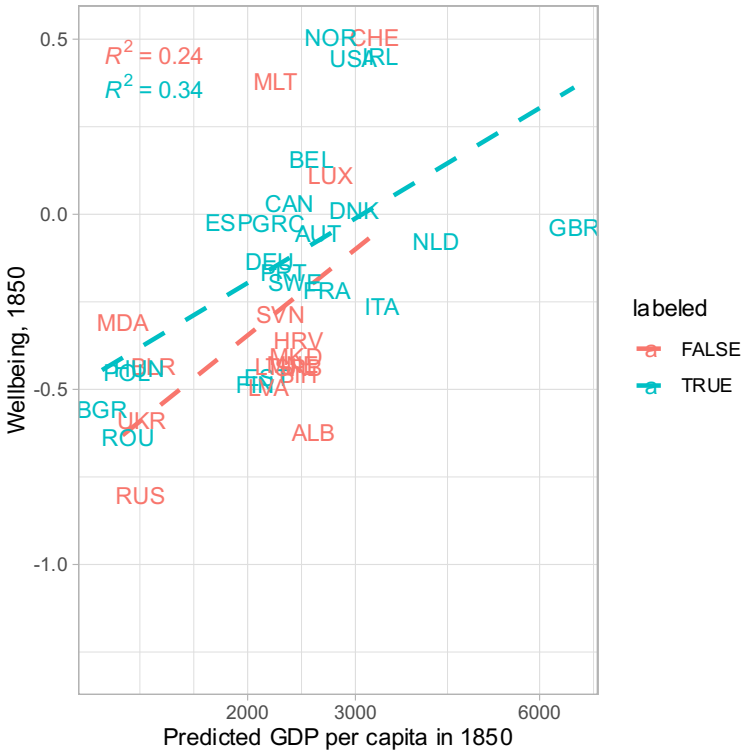



*Figure S29. Correlation between estimated GDP per capita and the OECD wellbeing indicator, for labeled (TRUE) and unlabeled (FALSE) observations*

## 5.4. German regions after the French Revolution

We explore whether our estimates replicate results by Acemoglu and coauthors regarding the economic development of German regions after the French Revolution (39).

Following the replication package the authors provide, we identify the NUTS-2 regions the 19 cities they are describing are in. Then, we compare the development of the treated group and control group the authors define in Table 1.

- Treated group: "DEB1", "DEB3", "DEA5", "DEA3", "DE91", "DEE0", "DE73", "DE92"
- Control group: "DE12", "DE21", "DE71", "DED2", "DE11", "DE40", "RUS.21_1", "PL42", "PL51", "DE80", "DEF0"

We use the population-weighted average to aggregate these groups, using the number of births and deaths of famous individuals as a proxy for population.

Our results (Figure S30) strongly resemble theirs using urbanization as a proxy for economic development (see their Fig. 2B for comparison).

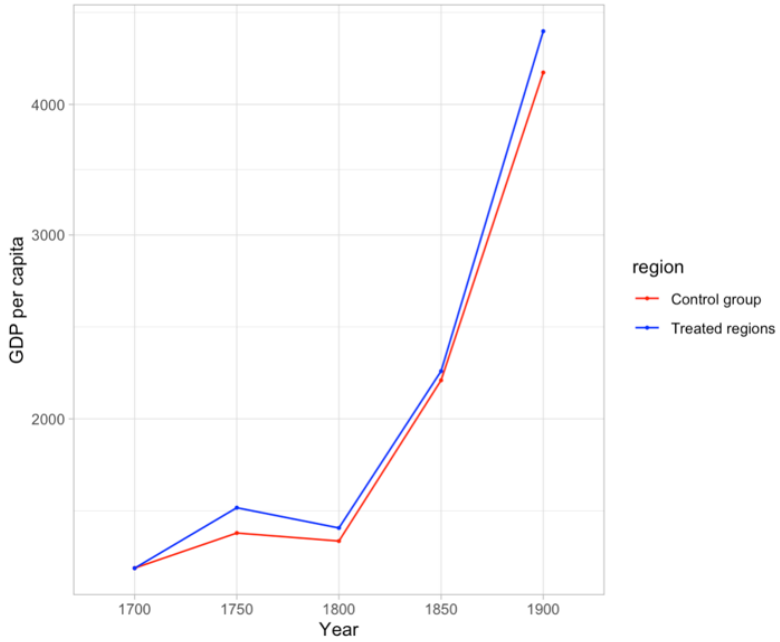



*Figure S30. Economic development in German regions occupied by the French army (treated) and other German regions (control)*

## 5.5. Robustness

### 5.5.1. Using only data prior to the year 2000

Our model performance results might be driven by the fact that our model is significantly better at predicting GDP per capita levels in the year 2000 than for other periods. Here, we exclude all observations of the year 2000 and rerun our model.

Indeed, model performance in terms of R-squared goes down. But the model performance in terms of the mean absolute error does not decrease.

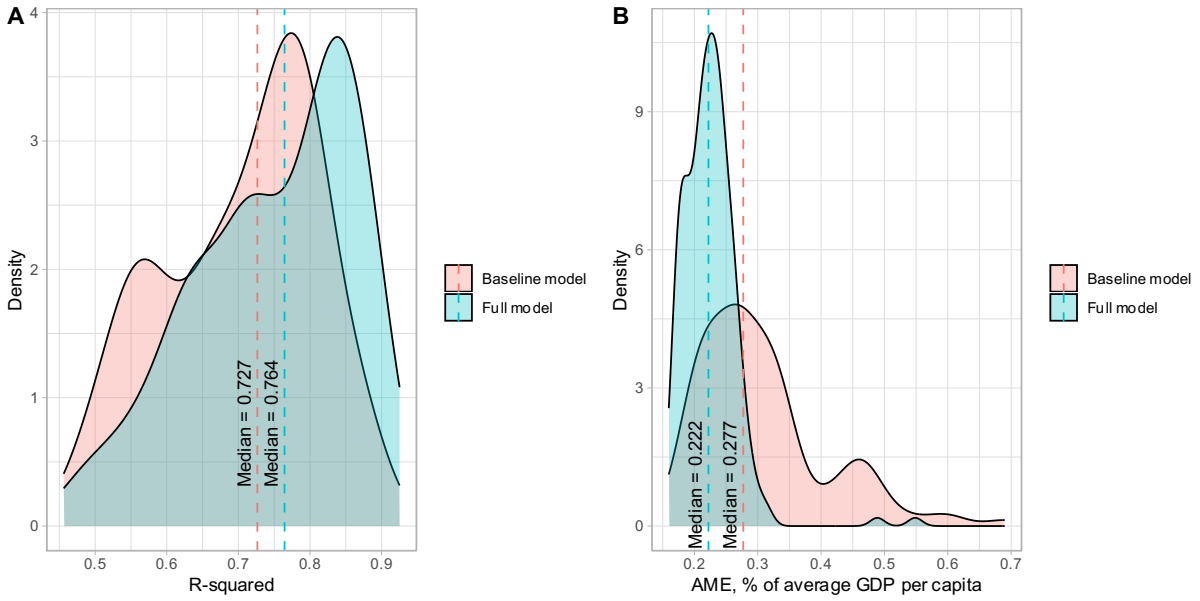



*Figure S31. Model performance observations prior to the year 2000.*

### 5.5.2. Comparing results across language editions

Wikipedia is known to have several biases, as discussed in the main manuscript. This includes an English bias. Since English Wikipedia is more comprehensive, individuals living in English speaking countries might be overrepresented. We address this issue by reducing our sample to biographies with Wikipedia pages in at least two language editions. This should reduce noise and limit the overrepresentation of biographies in English-speaking countries.

Also, we check whether our estimates are subject to an English bias by comparing our results to two other approaches of defining the sample: (1) using only pages that exist in English, and (2) using only non-English pages. If our estimates are prone to an English bias, we should observe substantial differences between the three estimates.

This, however, is not the case. Figure S32 shows the correlation of estimates obtained with these three samples of individuals. They are highly similar with a correlation coefficient of at

least 0.978. We would expect the largest differences comparing estimates for English-speaking countries. Figure S33 compares estimates for US regions based on the three samples and shows that they are highly similar as well, with a correlation coefficient of at least 0.951.

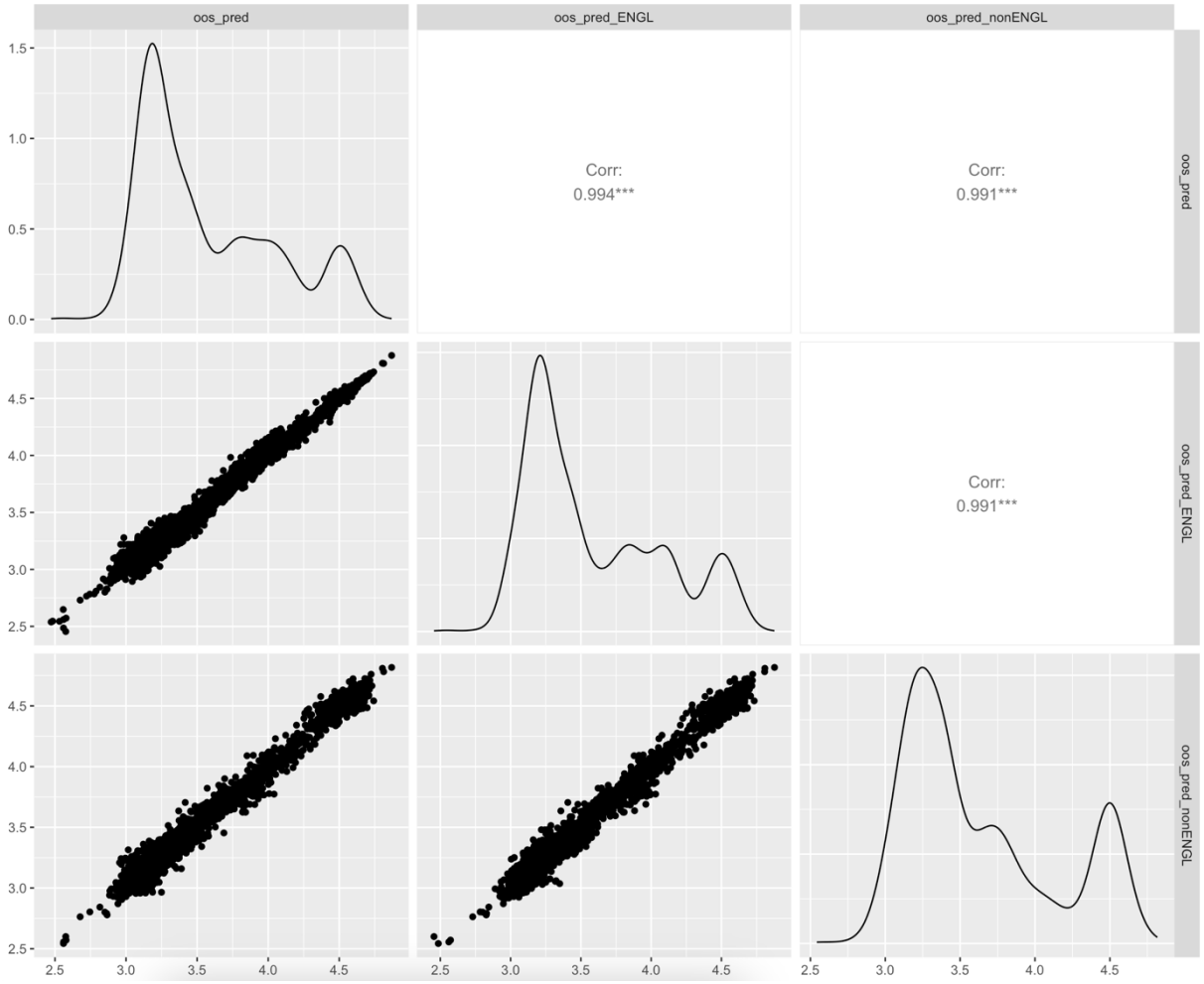



*Figure S32. Comparison between estimates using biographies with at least two language editions, only English pages, and only non-English pages.*

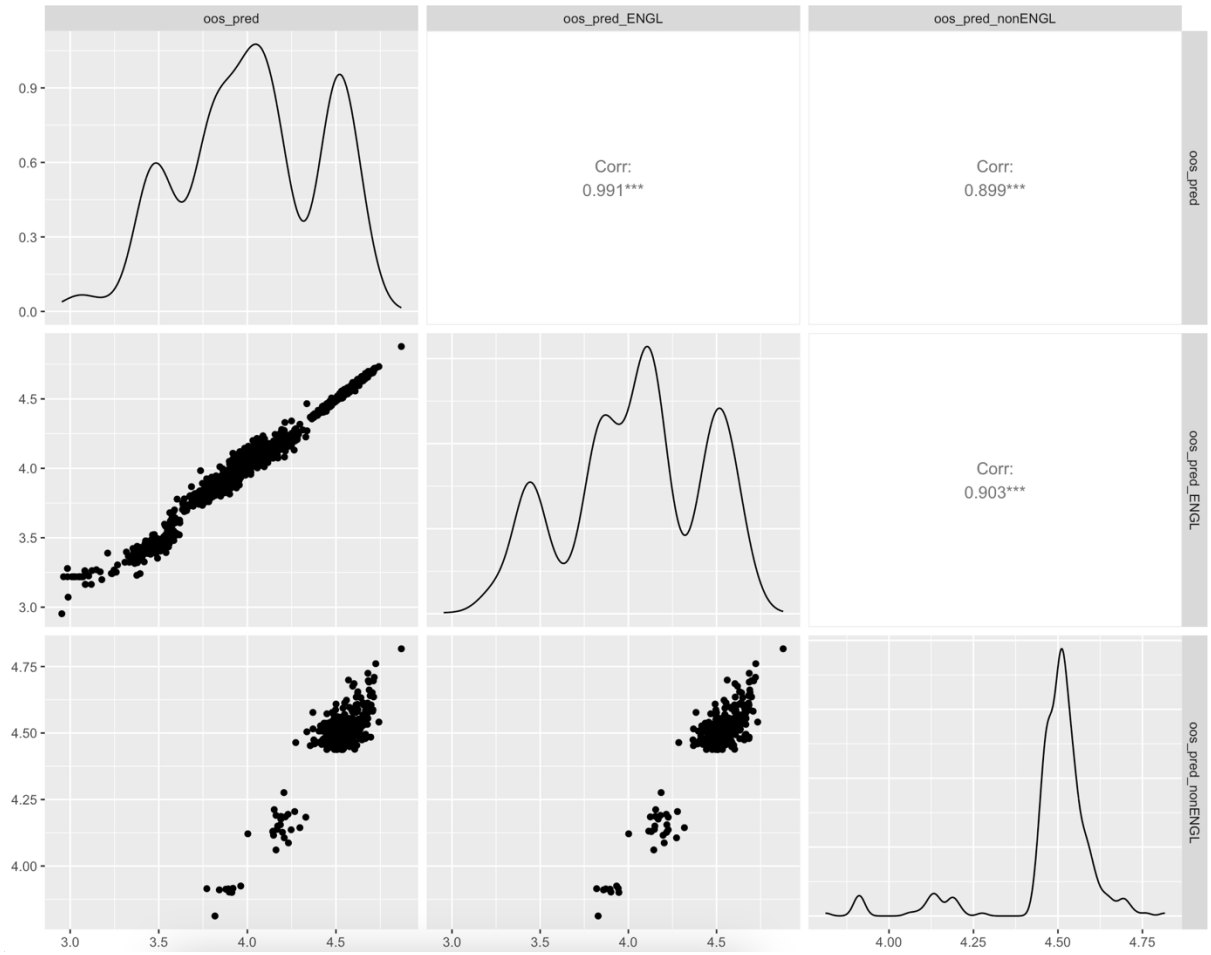



*Figure S33. Comparison of estimates for US metro- and micropolitan areas using biographies with at least two language editions, only English pages, and only non-English pages.*

### 5.5.3. Assignment of biographies to time periods

We assign individuals to time periods when they are born 150 prior to a certain year. Here, we test whether other threshold values yield different results. Specifically, we are investigating the thresholds 75 years (Figure S34), 100 years (Figure S35), and 175 years (Figure S36).

While all thresholds lead to an improvement in the out-of-sample estimates, none yields better results in terms of $R^2$ and absolute mean error than the model using 150 years as threshold (Fig. 2C-D in the main manuscript).

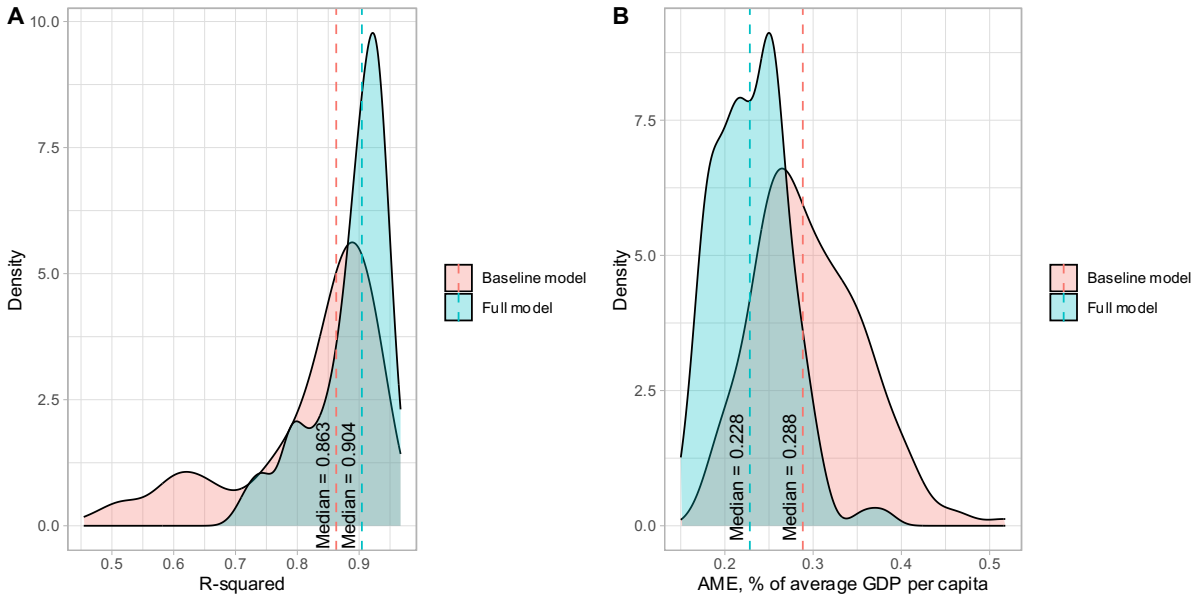



*Figure S34. Model performance using individuals born 75 years prior to a certain date for extracting features.*

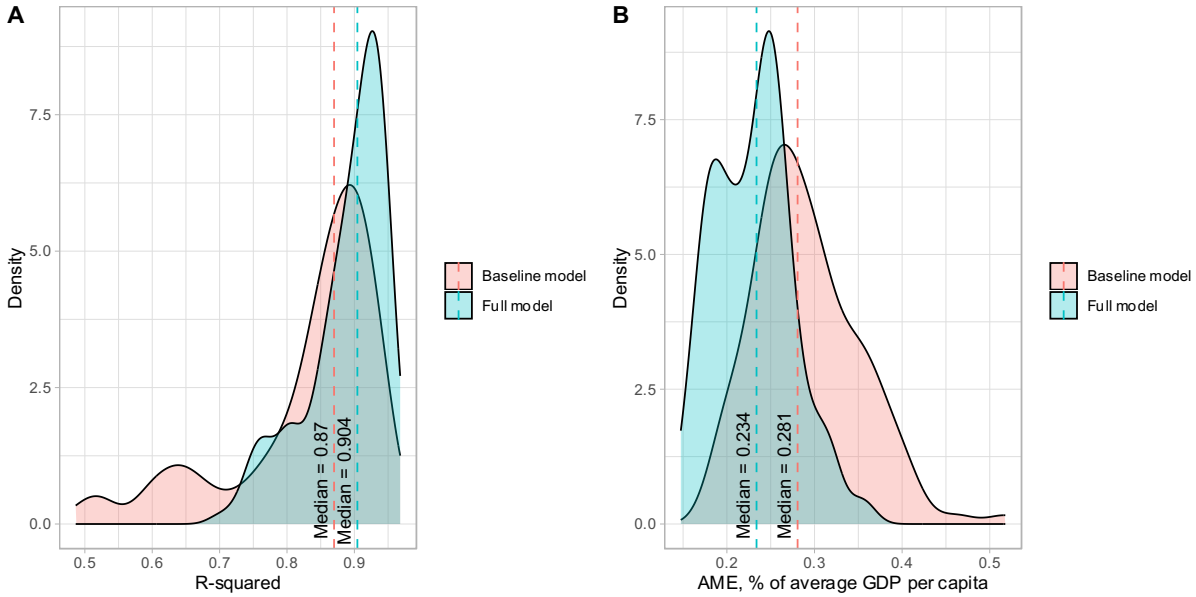



*Figure S35. Model performance using individuals born 100 years prior to a certain date for extracting features.*

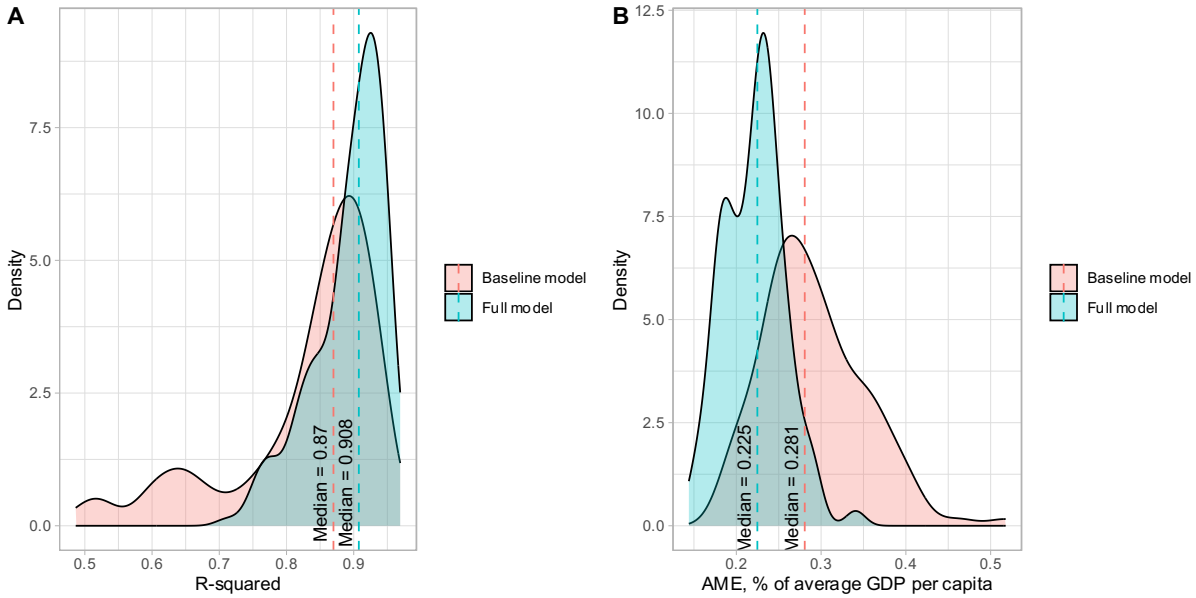



*Figure S36. Model performance using individuals born 175 years prior to a certain date for extracting features.*

### 5.5.4. Scaling features using the inverse hyperbolic sine function

We are scaling our features using logarithms. Specifically, we are using the function $\log(1 + x)$ to incorporate zeros. The inverse hyperbolic sine function is another approach that serves the same purpose. To show that our results are independent from our choice of the scaling function, we run our model using the inverse hyperbolic sine function. The model performance is very similar to using logarithms. While the R-squared is slightly better, the mean absolute error of the predictions is slightly worse (Figure S37).

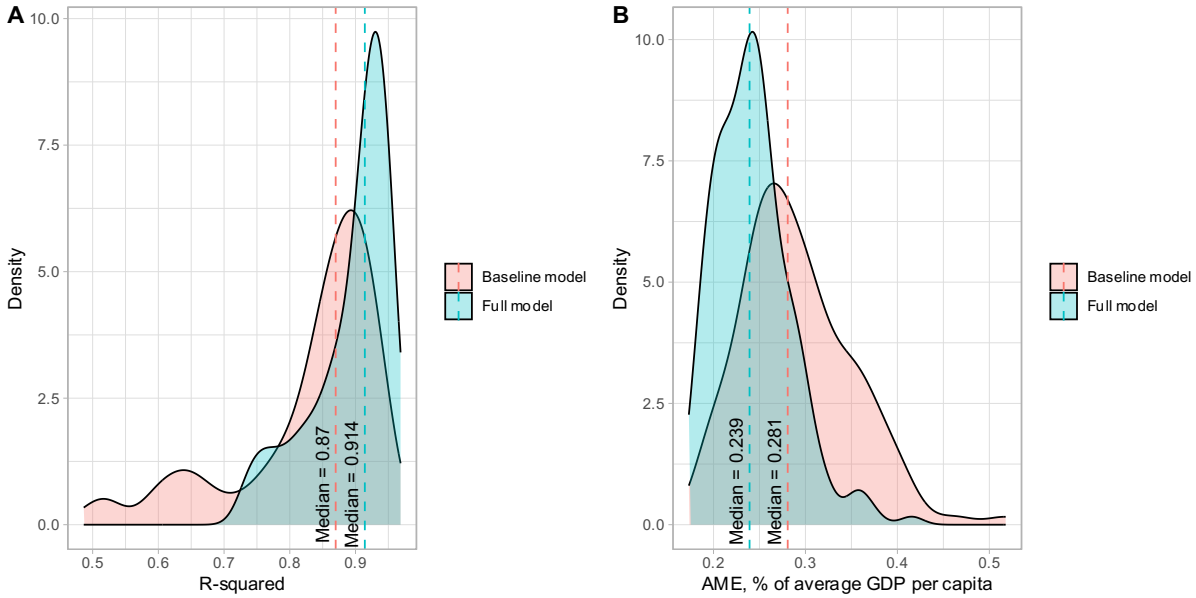



*Figure S37. Model performance using inverse hyperbolic sine function to scale features.*

### 5.5.5. Backward feature selection

We use a regularization technique, i.e. elastic net models, to select relevant features. Wrapper methods, such as backward feature selection, are also frequently used for this purpose. Here, we assess the model performance if using backward feature selection.

Backward feature selection works by recursively training the model with different subsets of features. Initially, all features are considered, and the model's performance is evaluated using k-fold cross validation. The least important feature (with respect to a feature's predictive power) is then eliminated, and the model is trained again with the reduced feature set. This process is repeated iteratively.

While we do not (need to) assume any fixed variables in the elastic net model, backward feature selection provided highly inaccurate results if no fixed variables were provided. Hence, we use the backward feature selection on top of our naïve baseline model. That is, we do not train the model with GDP per capita estimates, but with the residuals of regressing GDP per capita estimates on our baseline model. Even with this assumption, backward feature selection does not beat the baseline model in predicting the outcomes of independent test data sets with respect to the R-squared. Also, the model performance regarding the mean absolute error is lower (Figure S38).

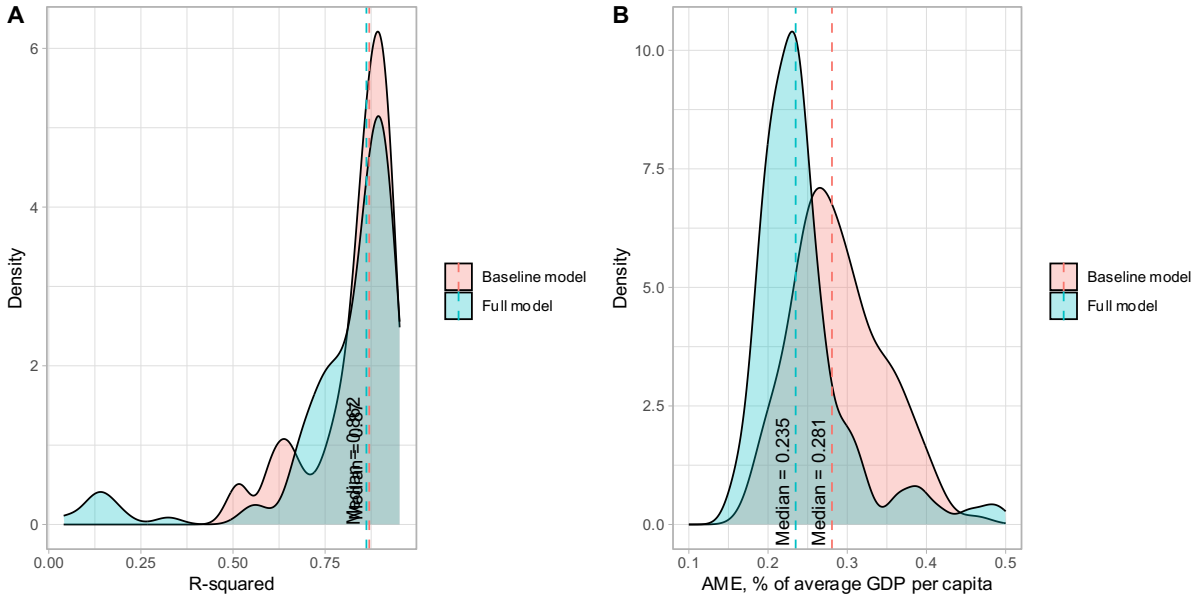



*Figure S38. Model performance using backward feature selection*

### 5.5.6. Using historical popularity to define features

In our main results, we use the Historical Popularity Index (HPI) as weights when defining features. Not using the HPI yields very similar model performance metrics (Figure S39).

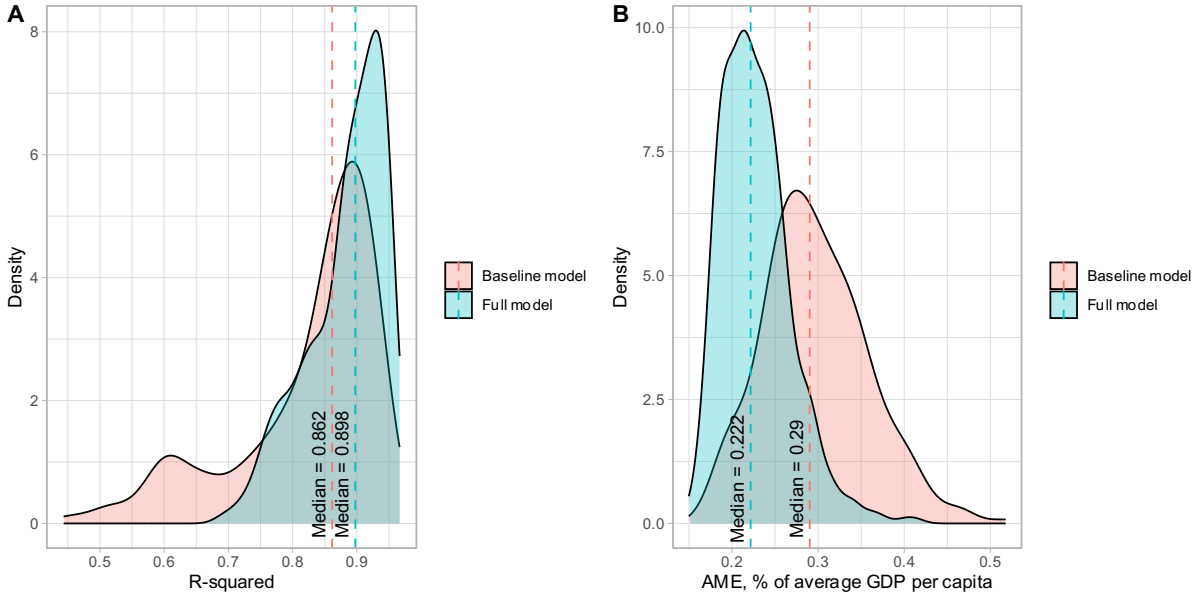



*Figure S39. Model performance when not using the HPI to weigh features.*

### 5.5.7. Removing dummies for supranational regions

We test whether the dummies for supranational regions are to a large extent driving our model performance results by removing them from the features the model can select. We find it provides highly similar results (Figure S40). These models still include the GDP per capita at the end of the previous period.

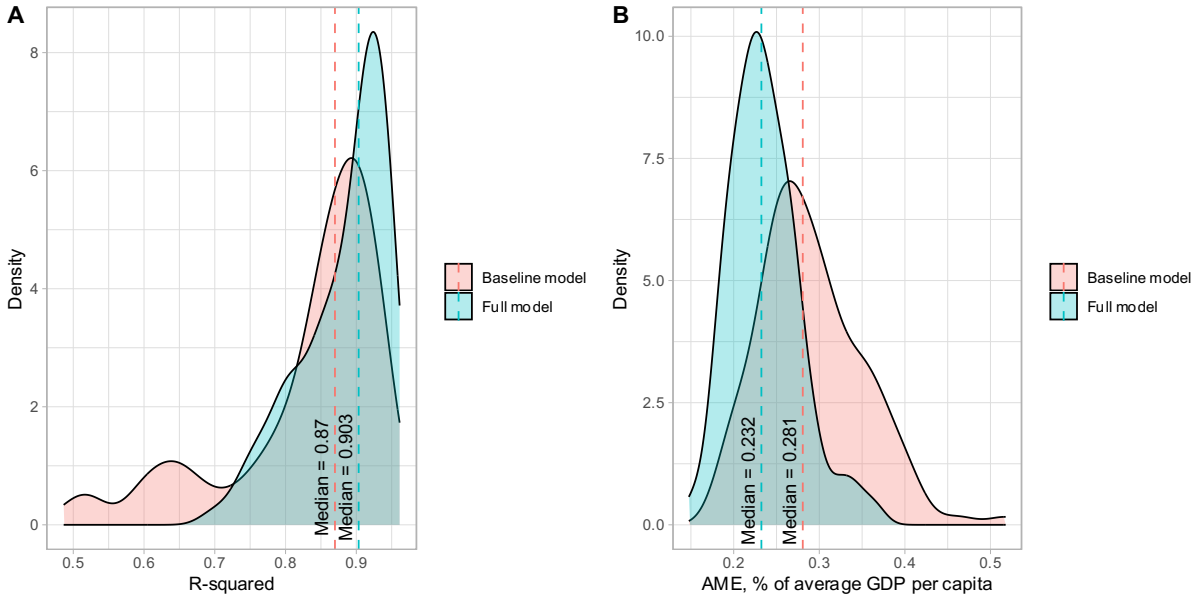



*Figure S40. Model performance when not using dummies for supranational regions.*

### 5.5.8. Predicting growth rates

An alternative to predicting GDP per capita levels is predicting GDP growth rates. We follow the same model setup to do so, but do not find positive results. Model performance metrics are significantly lower than when predicting GDP per capita levels, and there is no significant difference between the baseline model and the full model (see Figure S41).

We believe this is the case for two reasons. First, it is significantly harder to predict growth rates instead of levels. Second, we have a significantly lower amount of labeled training data. Specifically, we only have 455 true observations when predicting growth rates, while we can train our model on more than 1,300 observations when predicting levels.

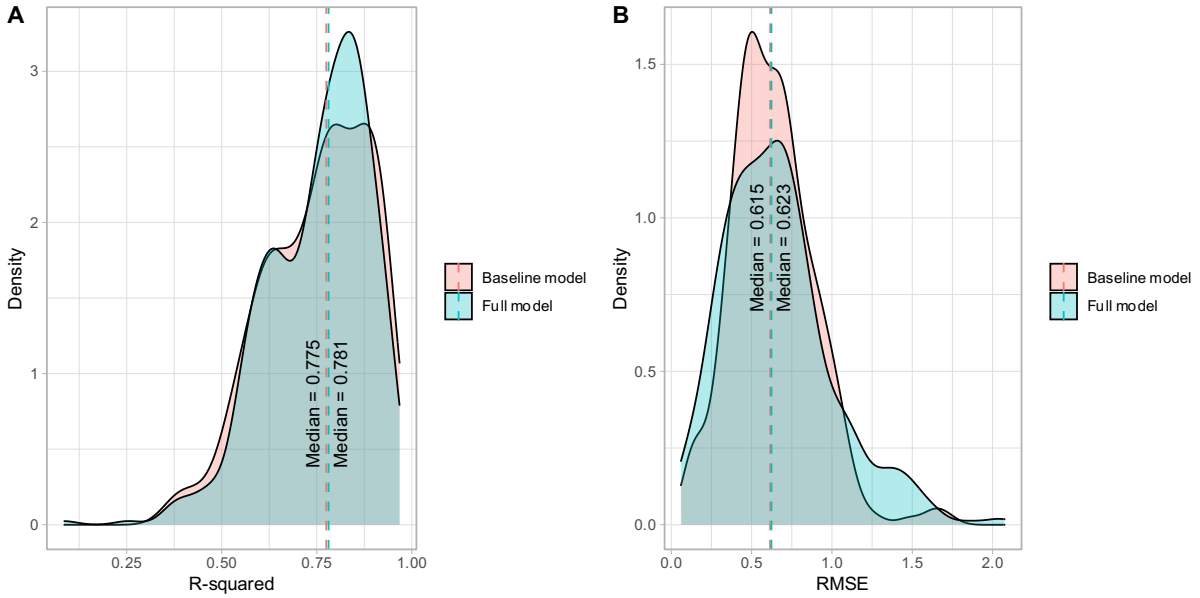



*Figure S41. Model performance when predicting growth.*

# 6. References


1. J. Bolt, J. L. van Zanden, Maddison style estimates of the evolution of the world economy. A new 2020 update. *Maddison-Proj. Work. Pap.* **WP-15** (2020).
2. J. Bolt, J. L. van Zanden, The Maddison Project: collaborative research on historical national accounts: The Maddison Project. *Econ. Hist. Rev.* **67**, 627–651 (2014).
3. F. Geary, T. Stark, Regional GDP in the UK, 1861-1911: new estimates: Regional GDP. *Econ. Hist. Rev.* **68**, 123–144 (2015).
4. Office for National Statistics, Historical Regional GDP 1968 to 1970 and 1971 to 1996. Office for National Statistics. Deposited 2016.
5. K. Enflo, M. Henning, L. Schön, "Swedish Regional GDP 1855–2000: Estimations and General Trends in the Swedish Regional System" in *Research in Economic History*, (Emerald Group Publishing, 2014), pp. 47–89.
6. K. Enflo, A. Missiaia, Regional GDP estimates for Sweden, 1571–1850. *Hist. Methods J. Quant. Interdiscip. Hist.* **51**, 115–137 (2018).
7. N. Delefortrie, J. Morice, Les revenus départementaux en 1864 et en 1954. *Population* **15**, 721 (1960).
8. M. P. Squicciarini, N. Voigtländer, Human Capital and Industrialization: Evidence from the Age of Enlightenment *. *Q. J. Econ.* **130**, 1825–1883 (2015).
9. E. Felice, The roots of a dual equilibrium: GDP, productivity, and structural change in the Italian regions in the long run (1871–2011). *Eur. Rev. Econ. Hist.* (2018). https://doi.org/10.1093/ereh/hey018.
10. C. Alvarez-Nogal, L. P. De La Escosura, The decline of Spain (1500-1850): conjectural estimates. *Eur. Rev. Econ. Hist.* **11**, 319–366 (2007).
11. M. Badia-Miró, J. Guilera, P. Lains, Reconstruction of the Regional GDP of Portugal, 1890 - 1980. *UB Econ. - Work. Pap.* **12/280** (2012).
12. E. Buyst, Reversal of Fortune in a Small, Open Economy: Regional GDP in Belgium, 1896-2000. *SSRN Electron. J.* (2009). https://doi.org/10.2139/ssrn.1586762.
13. Eurostat, Gross domestic product (GDP) at current market prices by NUTS 2 regions. Eurostat. Deposited 2023.
14. Office for National Statistics, Regional gross domestic product: all ITL regions. Office for National Statistics. Deposited 2022.
15. Bureau of Economic Analysis, Gross Domestic Product by Metropolitan Area. Bureau of Economic Analysis. Deposited 2018.
16. Statistics Canada, Metropolitan gross domestic product. Statistics Canada. Deposited 2014.
17. State Statistics Services Ukraine, Валовий регіональний продукт. Deposited 2013.
18. Belstat, Gross regional product at current prices. Deposited 2023.
19. Rosstat, Gross Regional Product at current basic prices per capita (1998-2019). Deposited 2020.
20. J. L. van Zanden, B. van Leeuwen, Persistent but not consistent: The growth of national income in Holland 1347–1807. *Explor. Econ. Hist.* **49**, 119–130 (2012).
21. S. N. Broadberry, B. M. S. Campbell, A. Klein, M. Overton, B. van Leeuwen, *British economic growth, 1270-1870* (Cambridge University Press, 2015).
22. N. Palma, J. Reis, From Convergence to Divergence: Portuguese Economic Growth, 1527–1850. *J. Econ. Hist.* **79**, 477–506 (2019).
23. P. Malanima, The long decline of a leading economy: GDP in central and northern Italy, 1300-1913. *Eur. Rev. Econ. Hist.* **15**, 169–219 (2011).

24. U. Pfister, Economic Growth in Germany, 1500–1850. *J. Econ. Hist.* **82**, 1071–1107 (2022).
25. M. Malinowski, J. L. van Zanden, Income and its distribution in preindustrial Poland. *Cliometrica* **11**, 375–404 (2017).
26. C. Álvarez-Nogal, L. P. De La Escosura, The rise and fall of Spain (1270-1850). *Econ. Hist. Rev.* **66**, 1–37 (2013).
27. O. Krantz, Swedish GDP 1300-1560 : A Tentative Estimate. *Lund Pap. Econ. Hist. Gen. Issues* **152** (2017).
28. L. Ridolfi, Six Centuries of Real Wages in France from Louis IX to Napoleon III: 1250–1860. *J. Econ. Hist.* **79**, 589–627 (2019).
29. L. Schön, O. Krantz, New Swedish Historical National Accounts since the 16th Century in Constant and Current Prices. *Lund Pap. Econ. Hist. Gen. Issues* **140** (2015).
30. M. Laouenan, *et al.*, A cross-verified database of notable people, 3500BC-2018AD. *Sci. Data* **9**, 290 (2022).
31. A. Z. Yu, S. Ronen, K. Hu, T. Lu, C. A. Hidalgo, Pantheon 1.0, a manually verified dataset of globally famous biographies. *Sci. Data* **3**, 150075 (2016).
32. P. Bairoch, J. Batou, P. Chèvre, *La population des villes européennes de 800 à 1850* (Librairie Droz, 1988).
33. E. Buringh, The Population of European Cities from 700 to 2000: Social and Economic History. *Res. Data J. Humanit. Soc. Sci.* **6**, 1–18 (2021).
34. M. Schich, *et al.*, A network framework of cultural history. *Science* **345**, 558–562 (2014).
35. M. Serafinelli, G. Tabellini, Creativity over time and space: A historical analysis of European cities. *J. Econ. Growth* **27**, 1–43 (2022).
36. P. Koch, V. Stojkoski, C. A. Hidalgo, The Role of Immigrants, Emigrants, and Locals in the Historical Formation of European Knowledge Agglomerations. *Reg. Stud.* (2023). https://doi.org/10.1080/00343404.2023.2275571.
37. C. A. Hidalgo, Economic complexity theory and applications. *Nat. Rev. Phys.* **3**, 92–113 (2021).
38. D. Acemoglu, S. Johnson, J. Robinson, The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth. *Am. Econ. Rev.* **95**, 546–579 (2005).
39. D. Acemoglu, D. Cantoni, S. Johnson, J. A. Robinson, The Consequences of Radical Reform: The French Revolution. *Am. Econ. Rev.* **101**, 3286–3307 (2011).